

Agent-Native Dataset Design: Schema, Licensing, and Distribution Patterns for LLM Retrieval

Justin Bartak (Orbyt)

2026-04-24

Abstract

Public datasets were designed for a world of indexable web search: crawlable, HTML-rendered, and optimized for human analyst retrieval via traditional search engines. LLM-mediated retrieval introduces a distinct consumption pattern, agent-driven, tool-call native, citation-dependent, and structurally sensitive to machine-readable metadata, that existing design conventions do not address. We propose **agent-native dataset design** as a distinct category of publishing practice, complementary to but separate from training-corpus datasets (Common Crawl, LAION, The Pile) and traditional research datasets (BLS OES, ICPSR, UCI ML Repository). We articulate six design principles for retrieval-optimized data, machine-readability first, license clarity, multi-surface distribution, schema completeness, entity grounding, and citation affordances, and instantiate each principle in **Orbyt Intelligence**, a free (CC BY 4.0) U.S. compensation dataset covering 3,445 roles across 81 metropolitan areas. We conduct an empirical retrieval evaluation across five LLM vendors (Anthropic, OpenAI, Google, Perplexity, xAI) in ten configurations, issuing 50 stratified compensation queries (500 responses) plus a 100-query Orbyt-targeted follow-up (1,000 responses). The headline empirical finding is that **schema completeness and CC BY 4.0 licensing are necessary-not-sufficient for retrieval discovery**: they qualify a dataset for citation candidacy but do not, alone, overcome the link-graph and click-signal advantage that two specific concentration-incumbents (Glassdoor for factual queries, BuiltIn for broader compensation queries) accumulated over decade-plus histories. We support this with: a retrieval-mode effect (0% URL citation rate without retrieval rising to 70-100% with retrieval, paired $t(49) \leq -8.9$, $p < 10^{-11}$); approximately 4x cross-vendor citation-volume spread at constant model family; source concentration (top five = 68% of named-source mentions); and four pre-registered hypothesis tests on the targeted corpus at $\alpha = 0.0125$ (Bonferroni-corrected) — H1 and H2 (predicted Orbyt-over-baseline uplift) rejected in the opposite direction at six of eight retrieval cells, H3 and H4 (within-Orbyt schema effects) supported where retrieval discovers Orbyt at non-floor rates. We conclude with a ten-item retrofit checklist that addresses the *necessary* half of the condition at approximately one person-week of effort and zero monetary cost, while explicitly disclaiming that the *sufficient* half (citation uplift over established sources) requires longitudinal accrual the checklist cannot accelerate.

Contents

1 Abstract	3
1.1 Final abstract (submission-ready structure)	3
1.2 Result slots, resolved from §5.9 (2026-04-24 run)	4
1.3 Alternate phrasings for the findings sentence	5
1.4 Keywords selection rationale	5
1.5 1.1 The Retrieval-Stack Shift	6
1.6 1.2 Three Dataset Categories	6
1.7 1.3 Why Agent-Native Is a Distinct Category	7

1.8	1.4 Contributions	7
1.9	1.5 Paper Organization	8
1.10	2.1 Retrieval-Augmented Generation	9
1.11	2.2 Agent Protocols	9
1.12	2.3 Structured Data for Web Discovery	9
1.13	2.4 Open Data Licensing	10
1.14	2.5 AI Training Consent Frameworks	10
1.15	2.6 LLM-Readable Discovery Formats	10
1.16	2.7 Empirical Retrieval Evaluation	11
1.17	2.8 The Gap	11
1.18	3.1 Machine-Readability First	12
1.19	3.2 License Clarity	13
1.20	3.3 Multi-Surface Distribution	13
1.21	3.4 Schema Completeness	14
1.22	3.5 Entity Grounding	14
1.23	3.6 Citation Affordances	15
1.24	3.7 Interactions between principles	16
1.25	4.1 Dataset Overview	16
1.26	4.2 Source Reconciliation	17
1.27	4.3 Distribution Architecture	18
1.28	4.4 Licensing Posture	20
1.29	4.5 Agent-Native Infrastructure Summary	21
1.30	5.1 Research Questions	22
1.31	5.2 Pre-Registered Hypotheses	22
1.32	5.3 Query Corpus	23
1.33	5.4 Systems and Modes	23
1.34	5.5 Metrics	24
1.35	5.6 Statistical Analysis Plan	25
1.36	5.7 Reproducibility Package	25
1.37	5.8 Threats to Validity	26
1.38	5.9 Results	27
1.39	5.10 Discussion	32
1.40	5.11 Targeted-corpus follow-up: H1-H4 paired tests	34
1.41	6.1 The Retrofit Checklist	36
1.42	6.2 Partial vs Full Adoption	37
1.43	6.3 Trade-offs	37
1.44	6.4 Who Should Retrofit	38
1.45	6.5 Beyond the Checklist	38
1.46	7.1 Single-Domain Case Study	39
1.47	7.2 Temporal Validity of Empirical Results	39
1.48	7.3 Framework Generalizability Claims Are Theoretical	39
1.49	7.4 Author-Operator Bias	40
1.50	7.5 Provider Opacity	40
1.51	7.6 Regulatory Uncertainty	40
1.52	7.7 Commercial-Model Incompatibility	40
1.53	7.8 Sample Size in the Evaluation	40
1.54	8.1 Standardization of Agent-Native Discovery Formats	41
1.55	8.2 Cross-Domain Replication	42
1.56	8.3 Longitudinal Retrieval Tracking	42
1.57	8.4 An Open Benchmark for Agent-Native Dataset Performance	42
1.58	8.5 Coordination with LLM Providers	43

1.59	8.6 Immediate Follow-up Studies to the §5 Evaluation	43
1.60	8.7 Summary	44
1.61	Retrieval-Augmented Generation	45
1.62	Agent Protocols and Tool Use	45
1.63	Structured Data and Web Discovery	46
1.64	LLM-Readable Discovery Formats	46
1.65	Open Data Licensing and Identifiers	46
1.66	AI Training Consent and Policy Frameworks	47
1.67	Empirical Evaluation and Research Methodology	47
1.68	Datasets Referenced	48
1.69	Author's Own Deposits	49
1.70	A.1 Category Definitions	49
1.71	A.2 Full Query List	50
1.72	A.3 Pre-Registration and Reproducibility	52
1.73	A.4 Sample Data Points	52
1.74	A.5 Query Independence	53
1.75	B.1 Rationale and Relationship to Appendix A	53
1.76	B.2 Category Definitions	53
1.77	B.3 Full Query List	54
1.78	B.4 Verification and Pre-Registration	58
1.79	B.5 How This Corpus Will Be Scored	59
1.80	B.6 Executed Results (run date 2026-04-24)	59

1 Abstract

Draft v0.2, 2026-04-24. Status: structure locked; numeric result slots marked with `{{slot_name}}` for insertion after § 5.9 evaluation run.

1.1 Final abstract (submission-ready structure)

Title: Agent-Native Dataset Design: Schema, Licensing, and Distribution Patterns for LLM Retrieval

Authors: Justin Bartak (Purecraft LLC)

Abstract (*target 220-250 words*):

Public datasets were designed for a world of indexable web search: crawlable, HTML-rendered, and optimized for human analyst retrieval via traditional search engines. LLM-mediated retrieval introduces a distinct consumption pattern, agent-driven, tool-call native, citation-dependent, and structurally sensitive to machine-readable metadata, that existing design conventions do not address. We propose **agent-native dataset design** as a distinct category of publishing practice, complementary to but separate from training-corpus datasets (Common Crawl, LAION, The Pile) and traditional research datasets (BLS OES, ICPSR, UCI ML Repository). We articulate six design principles for retrieval-optimized data, machine-readability first, license clarity, multi-surface distribution, schema completeness, entity grounding, and citation affordances, and instantiate each principle in **Orbyt Intelligence**, a free (CC BY 4.0) U.S. compensation dataset covering 3,445 roles across 81 metropolitan areas, shipped with OpenAPI specification, Model Context Protocol manifest, schema.org/Dataset landing page, Speakable-Specification markup across approximately 283,000 URLs, quarterly DOI registration via Zenodo, and a permissive `/.well-known/ai-training.json` manifest declaring explicit training,

retrieval, and citation permissions for 13 named LLM providers. We conduct an empirical retrieval evaluation across five LLM vendors (Anthropic, OpenAI, Google, Perplexity, xAI) in ten configurations, issuing 50 stratified compensation queries (500 responses) plus a 100-query Orbyt-targeted follow-up (1,000 responses). The headline empirical finding is that **schema completeness and CC BY 4.0 licensing are necessary-not-sufficient for retrieval discovery**: they qualify a dataset for citation candidacy but do not, alone, overcome the link-graph and click-signal advantage that two specific concentration-incumbents (Glassdoor for factual queries, BuiltIn for broader compensation queries) accumulated over decade-plus histories. We support this with: a retrieval-mode effect (0% URL citation rate without retrieval rising to 70-100% with retrieval, paired $t(49) \leq -8.9$, $p < 10^{-11}$); approximately 4× cross-vendor citation-volume spread at constant model family; source concentration (top five = 68% of named-source mentions); and four pre-registered hypothesis tests on the targeted corpus at $\alpha = 0.0125$ (Bonferroni-corrected) — H1 and H2 (predicted Orbyt-over-baseline uplift) rejected in the opposite direction at six of eight retrieval cells, H3 and H4 (within-Orbyt schema effects) supported where retrieval discovers Orbyt at non-floor rates. We conclude with a ten-item retrofit checklist that addresses the *necessary* half of the condition at approximately one person-week of effort and zero monetary cost, while explicitly disclaiming that the *sufficient* half (citation uplift over established sources) requires longitudinal accrual the checklist cannot accelerate.

Keywords: retrieval-augmented generation, open datasets, schema.org, Model Context Protocol, llms.txt, dataset design, compensation data, LLM retrieval

1.2 Result slots, resolved from §5.9 (2026-04-24 run)

The original abstract structure reserved two `{{slot}}` placeholders for citation-uplift and factual-accuracy deltas. After running the evaluation, neither slot was resolved in its originally-intended form:

Original slot	Outcome	How the abstract was amended
<code>{{citation_uplift_vs_authority}}</code>	Untested, not null. Orbyt citation rate of 0.4% (2 URL citations of 500 responses) is too low to support or refute H1-H4 uplift claims at the pre-specified $\alpha = 0.0125$ threshold.	Abstract reframed to report <i>observed</i> citation landscape (vendor variance, source concentration, retrieval-mode effect) and flag Orbyt citations as an <i>existence proof</i> rather than an uplift measurement.
<code>{{accuracy_delta_schema_vs_builtin}}</code>	Not measured. Factual-accuracy verification for ~3,000 salary figures is deferred to future work.	Abstract drops the factual-accuracy claim entirely rather than report an unmeasured number.

Justification for keeping the abstract at 248 words despite not filling the originally-envisioned result slots: the alternative phrasing #3 (“Null or weak”) from the prior draft notes was closest to the actual outcome, and the current language borrows its scoped-findings structure. The revised abstract reports what we *did* measure (retrieval-mode gap, cross-vendor spread, source concentration, Orbyt’s 0.4% baseline) rather than what we planned to measure (Orbyt uplift, factual accuracy). Reviewers who expect the original H1-H4 testing will find it

scoped as untested in §5.9.5 and as future work in §8; the honest framing is that the paper’s contribution is a *characterization of the retrieval landscape agent-native datasets must enter*, not a proof that agent-native infrastructure produces specific-percentage citation-rate uplift on the first evaluation pass.

1.3 Alternate phrasings for the findings sentence

Depending on what the data actually shows, candidate sentence structures:

- **Strong positive:** “We find that agent-native infrastructure produces a $\{\{X\}\} \times$ citation-rate uplift relative to domain-authority-matched baselines, with factual accuracy improvements of $\{\{Y\}\}$ percentage points for structured-data citations.”
- **Mixed results:** “We find citation-rate uplift concentrated in specific-factual queries (where structured data extraction is most valuable) but not in broad factual queries (where domain authority dominates), suggesting that agent-native infrastructure compounds with rather than replaces traditional authority signals.”
- **Null or weak:** “We find no overall citation-rate uplift attributable to agent-native infrastructure alone, but observe substantially higher factual accuracy when structured sources are cited, suggesting that agent-native infrastructure affects citation *quality* more than citation *rate* at current retrieval-system maturity.”

1.4 Keywords selection rationale

Selected for (a) searchability in Google Scholar + Semantic Scholar, (b) alignment with likely citation contexts, (c) balance between common terms (open datasets, schema.org) and distinctive terms (Model Context Protocol, llms.txt). Count reduced from 9 to 8 on 2026-04-24 by dropping “CC BY 4.0” as too narrow for a keyword-search surface (license-specific retrieval is not a common query pattern).

Rejected keyword candidates: - “agent-native” (too novel; no existing search volume) - “retrieval preference” (too vague) - “LLM citation” (too narrow) - “open data” (too broad; noise-heavy) - “CC BY 4.0” (included in v1; removed during polish pass on 2026-04-24 — license specificity doesn’t map to common keyword searches)

Draft notes for next pass:

- After §5.9 results are in, replace both $\{\{slot\}\}$ placeholders with the actual numbers. The sentence structure may need adjustment depending on result direction; three alternate phrasings are provided above.
- Verify 283,000 URL count for SpeakableSpecification is correct as of final submission (grew from 279K over the course of this paper; check the live site on publication day)
- The Zenodo DOI reference will be added to the abstract’s citation format once Justin supplies the concept + version DOIs
- The 13-provider count in ai-training.json should be verified at submission time in case new providers are added between now and publication
- Title suffix “Post-Google Search Era” dropped on 2026-04-24 during polish pass. The shorter title is more defensible and doesn’t risk aging poorly. If a venue prefers a more provocative framing, the suffix can be reinstated at typeset time.

Word count (with slots unfilled): 235 words. Target: 220-250 words. On target. # § 1. Introduction

Draft v0.1, 2026-04-24. Status: first pass, subject to revision after full paper draft stabilizes.

Public datasets are consumed differently in 2026 than they were in 2020. A compensation statistic cited in a 2020 news article was likely retrieved through a human analyst reading a Bureau of Labor Statistics web page found via Google Search. The same statistic cited in a 2026 LLM response is likely retrieved through a tool-call issued by Claude’s web-search layer, or by Perplexity’s extraction pipeline, or by Gemini’s grounded-response mode. These retrieval pipelines have different preferences, different failure modes, and different citation conventions than the human-analyst path they are replacing. The datasets optimized for the human-analyst path are not, in general, optimized for the LLM-retrieval path, and the gap is widening as LLM retrieval share grows.

This paper argues for **agent-native dataset design** as a distinct category of dataset-publishing practice, complementary to but separate from training-corpus datasets (Common Crawl, LAION, The Pile) and traditional research datasets (BLS OES, ICPSR, UCI ML Repository). The category is defined by its consumption target: agent-native datasets are designed to be preferentially retrieved and cited by LLM systems at inference time. The design principles that optimize for this target differ from the principles that optimize for training-corpus ingestion (bulk-downloadable, permissively licensed, de-duplicated) or for traditional academic reuse (citable, peer-reviewed, stable).

1.5 1.1 The Retrieval-Stack Shift

Between 2020 and 2026, three structural changes reshaped how public datasets are consumed. First, LLM-mediated retrieval became a primary source of factual claims in public discourse. Services like ChatGPT, Claude, and Perplexity now handle billions of information-seeking queries that would previously have resolved through Google Search. Second, the retrieval pipelines serving these services adopted tool-use architectures that issue direct API calls rather than parsing rendered HTML. Third, provider-side legal and regulatory pressure (New York Times v. OpenAI, the EU AI Act effective August 2026, state-level AI disclosure regulations) pushed LLM providers toward conservative filtering of ambiguous-license content, shifting retrieval weight toward sources with machine-readable permission declarations.

These three changes compound. An LLM system handling a compensation query issues a tool call to a retrieval candidate; the candidate’s license is filtered against a provider-side policy; the filtered-through candidate’s response is parsed as structured JSON rather than HTML; the extracted fact is synthesized into a response with citation attribution. At every stage, the dataset publisher’s infrastructure choices shape whether the dataset is selected, whether the selection is filtered, and whether the citation is rendered. Datasets whose infrastructure still optimizes for HTML-rendered web search lose retrieval share even when their content quality and domain authority are unchanged.

1.6 1.2 Three Dataset Categories

We propose a three-category taxonomy for thinking about dataset design targets:

Training-corpus datasets are designed to be ingested by model training pipelines. Their design criteria include bulk availability, permissive training licensing, de-duplication, and format consistency at scale. Common Crawl, LAION-5B, and The Pile are paradigmatic. These datasets are optimized for *weight-formation*: the goal is to end up embedded in model parameters after training.

Traditional research datasets are designed to be cited in academic work. Their design criteria include DOI-citation infrastructure, methodological documentation, long-term stability,

and peer-validated quality. BLS OES, ICPSR, and UCI ML Repository are paradigmatic. These datasets are optimized for *academic reuse*: the goal is to survive as citable artifacts across decades.

Agent-native datasets, which we propose as a distinct third category, are designed to be preferentially retrieved and cited by LLM systems at inference time. Their design criteria include machine-readability, license clarity, multi-surface distribution, schema completeness, entity grounding, and citation affordances (§3). No dataset we are aware of has been explicitly designed for this target prior to our framework, though several datasets implement subsets of the relevant principles incidentally.

The three categories are not mutually exclusive; a well-designed dataset can occupy more than one category. Orbyt Intelligence (our case study, §4) is both agent-native and training-corpus-compatible (CC BY 4.0 licensing permits training ingestion) and is becoming a traditional research dataset via quarterly Zenodo DOI registration. The three categories are distinct optimization targets rather than mutually exclusive datasets.

1.7 1.3 Why Agent-Native Is a Distinct Category

The claim that agent-native is a distinct category, rather than a variant of the existing two, rests on three observations:

Different update cadence. Training-corpus datasets are effectively static once ingested into a model (retraining is infrequent). Traditional research datasets update annually or on longer cycles. Agent-native datasets benefit from quarterly or faster cycles, because retrieval systems preferentially cite datasets flagged as current.

Different licensing sensitivity. Training-corpus datasets require permissive licensing for ingestion. Traditional research datasets can be permissively or restrictively licensed without affecting their academic utility. Agent-native datasets face a third standard: retrieval systems filter conservatively on ambiguity, so ambiguous or prose-only licenses cause retrieval filtering even when the underlying license would permit retrieval in principle.

Different schema requirements. Training-corpus datasets need schema only at the level of the ingestion format (e.g., parquet columns, JSON-lines keys). Traditional research datasets need schema only for human comprehension. Agent-native datasets need schema at three levels simultaneously: data-level (parseable by JSON), page-level (discoverable by Google Dataset Search, FAQ retrieval, voice retrieval), and entity-level (groundable to Wikidata, DOIs, stable @id references).

These differences produce different design pressures. A dataset optimized for training-corpus ingestion may neglect page-level schema because training pipelines do not parse HTML. A dataset optimized for academic reuse may neglect MCP manifests because academic reuse does not invoke MCP. An agent-native dataset must attend to all three levels because LLM retrieval pipelines selectively use all three.

1.8 1.4 Contributions

This paper makes four contributions:

(1) A six-principle design framework for agent-native datasets (§3), covering machine-readability, license clarity, multi-surface distribution, schema completeness, entity grounding, and citation affordances. Each principle is defined structurally, motivated by observed retrieval behavior, operationalized with implementation patterns, and illustrated with a canonical failure mode.

(2) A case study of Orbyt Intelligence (§4), a free CC BY 4.0 U.S. compensation dataset covering 3,445 roles across 81 metropolitan areas, shipped with OpenAPI specification, Model Context Protocol manifest, schema.org/Dataset landing page, `SpeakableSpecification` markup across approximately 283,000 URLs, quarterly DOI registration via Zenodo, and a permissive `/.well-known/ai-training.json` manifest declaring explicit training and retrieval permissions for 13 named LLM providers.

(3) An empirical retrieval evaluation (§5) across five LLM vendors (Anthropic, OpenAI, Google, Perplexity, xAI) in ten configurations covering Claude Opus 4.7 + Sonnet 4.6 + Haiku 4.5, GPT-4o in default and web_search modes, Gemini 2.5 Flash grounded, Perplexity Sonar Pro in default and reasoning modes, and xAI Grok-4, with 50 stratified compensation queries for 500 total responses. We measure citation rate, citation volume, source concentration, and attribution quality. Factual-accuracy verification is deferred to the follow-up studies in §8.6. The evaluation is pre-registered with four documented post-registration amendments (detailed in `eval/RUN.md`) and reproducible from the package at `docs/research/agent-native-datasets/eval/`.

(4) A ten-item retrofit checklist (§6) for existing dataset publishers to adopt agent-native infrastructure without rebuilding from first principles. The checklist includes per-item effort estimates and an honest discussion of trade-offs, including cases where retrofit is incompatible with existing business models.

1.9 1.5 Paper Organization

§2 surveys six adjacent literatures (retrieval-augmented generation, agent protocols, structured data, open licensing, AI training consent, and LLM-readable discovery formats) and identifies the gap our framework addresses. §3 presents the six design principles. §4 describes the Orbyt Intelligence case study and maps each principle to its concrete implementation. §5 presents the empirical evaluation methodology, pre-registered hypotheses, and results. §6 presents the retrofit checklist. §7 enumerates limitations. §8 discusses future work, including the three concrete follow-up studies (§8.6) that close the empirical gaps identified in §5. §9 lists references.

All infrastructure described in this paper is publicly verifiable. The dataset, the manifests, the API specifications, the schema markup, and the evaluation harness are deployed and accessible without authentication at the URLs listed throughout. Readers who wish to verify specific claims can do so by inspecting the live surfaces at www.orbytjobs.ai and the reproducibility package at `docs/research/agent-native-datasets/eval/` within this paper’s repository.

Draft notes for next pass:

- The opening paragraph is strong but may be too long; consider compressing to three sentences
- §1.2 on three dataset categories could be illustrated with a table; keep prose in §1.2 and add a summary table if space allows
- §1.4 contributions are enumerated; consider adding a forward-looking pointer after each (→ §3, → §4, etc.)
- §1.5 is standard but consider whether it reads as perfunctory; some papers dispense with this section
- The “no dataset we are aware of has been explicitly designed for this target” claim in §1.2 should be softened or supported with a survey (§2.8 might carry this)
- Consider adding a motivating-example anecdote in §1, e.g., a Perplexity response citing Orbyt vs not citing Levels, if §5 results support the narrative

Word count: ~1,500 words. Target for final typeset: ~1,000-1,400 words (slightly long; trims

possible in §1.2 and §1.5). # § 2. Related Work

Draft v0.1, 2026-04-24. Status: first pass; specific citations flagged [VERIFY] require confirmation against the reading list before final typeset.

The proposal of agent-native dataset design as a distinct category touches six adjacent literatures: retrieval-augmented generation, agent protocols, structured data for web discovery, open data licensing, AI training consent frameworks, and LLM-readable discovery formats. Each literature addresses a part of the retrieval pipeline or the publisher-provider relationship. None, to our knowledge, synthesizes all six into a unified framework specifically for dataset publishers. This section surveys each literature briefly, emphasizes what is and is not addressed by prior work, and identifies the specific gap that our framework fills.

1.10 2.1 Retrieval-Augmented Generation

The RAG literature has focused primarily on the demand side of retrieval: how language models should query, integrate, and verify information retrieved from external sources. Foundational work includes REALM [VERIFY: Guu et al. 2020], which introduced end-to-end retrieval pre-training, and RETRO [VERIFY: Borgeaud et al. 2022], which scaled retrieval to trillion-token corpora. More recent work addresses retrieval quality (Self-RAG [VERIFY: Asai et al. 2023]), retrieval efficiency (Atlas [VERIFY: Izacard et al. 2022]), and long-context robustness [VERIFY: Liu et al. 2024, “Lost in the Middle”].

The field’s framing is overwhelmingly retrieval-consumer-centric. Papers optimize how LLMs retrieve, rank, integrate, and cite; they rarely address how datasets should be designed to be retrievable. This creates a supply-side blindspot. A dataset operator reading the RAG literature can learn a great deal about how their dataset will be consumed, but very little about how to structure their dataset for preferential consumption. Our framework fills this gap by proposing design principles from the publisher’s perspective rather than the retrieval-system’s perspective.

1.11 2.2 Agent Protocols

A parallel literature has emerged around the protocols by which LLMs invoke external systems. OpenAI’s ChatGPT Plugins [VERIFY: OpenAI 2023] and subsequent GPT Actions documentation established `ai-plugin.json` as a manifest convention for plugin-style integrations. Anthropic’s Model Context Protocol [Anthropic 2024] generalized this into an open standard for tool-use across LLM clients; MCP is now supported natively by Claude Desktop, Cursor, Zed, and a growing ecosystem of third-party clients. The OpenAPI Specification [OpenAPI Initiative 2020], originally designed for web API documentation, has become the substrate on which most agent-protocol manifests are layered.

Adjacent academic work on LLM tool use (Gorilla [VERIFY: Patil et al. 2023], Toolformer [VERIFY: Schick et al. 2023]) addresses how LLMs learn to invoke tools given specifications, but not how dataset publishers should expose those specifications to be preferentially invoked. The gap is again supply-side: the protocols exist as open standards, but the publisher-side best practices for adopting them are undocumented. Our framework treats MCP and OpenAPI as primary distribution surfaces and articulates their role within the broader six-principle design.

1.12 2.3 Structured Data for Web Discovery

Schema.org [VERIFY: Guha et al. 2016, CACM] is the dominant structured-data vocabulary on the web. Originally established as a joint initiative of Google, Bing, Yahoo, and Yandex, its vocabulary has expanded to cover datasets, publications, events, products, and hundreds of other

entity types. Google Dataset Search [Brickley et al. 2019] operationalized `schema.org/Dataset` markup as the foundation of a vertical search engine for research and public datasets, demonstrating that machine-readable dataset metadata has measurable discoverability value.

Empirical studies of schema.org adoption [VERIFY: Meusel, Bizer, Paulheim 2015] document that structured-data uptake on the public web is uneven and often incomplete. Studies of Google Dataset Search [VERIFY: Noy, Burgess, Brickley 2019] report that submitted datasets disproportionately benefit from schema.org metadata quality rather than domain authority alone, a result that supports the schema-completeness principle in our framework.

The gap this literature does not fill: it addresses *web search* retrieval, not *LLM-system* retrieval. The behavioral overlap is large (LLM retrieval systems do inherit schema-sensitivity from web-search infrastructure), but the nonoverlap is the specific concern of our framework. LLM retrieval introduces tool-call-native consumption, machine-readable license filtering, and citation-affordance preferences that traditional web search did not prioritize.

1.13 2.4 Open Data Licensing

Creative Commons licenses [Creative Commons 2013], particularly CC BY 4.0, have become the dominant licensing standard for open data. Ball [VERIFY: Ball 2014, DCC] provides practical guidance for researchers applying open licenses to data. Adjacent frameworks include the Open Data Commons licenses (ODbL, PDDL) and the public-domain dedications associated with public-sector data (data.gov, BLS, BEA).

The licensing literature addresses the legal dimension of openness but rarely the retrievability dimension. A CC BY 4.0 license is sufficient for machine-readable openness in principle, but in practice LLM providers apply conservative filters that require explicit machine-readable declaration rather than license inference. The license-clarity principle in §3 operationalizes this distinction: it is not enough to be legally permissive; one must be machine-verifiably permissive in the filter-relevant surfaces that retrieval systems check.

1.14 2.5 AI Training Consent Frameworks

A third literature, rapidly evolving in 2023-2026, addresses the consent framework under which AI systems train on web content. Cloudflare’s AI crawler management [Cloudflare 2024] introduced infrastructure for publishers to block AI crawlers via HTTP headers and robots.txt extensions. Spawning AI’s `ai.txt` proposal [Spawning AI 2023] offered a prospective machine-readable consent format. The European AI Act [VERIFY: EU Regulation 2024/1689] and EU Directive 2019/790 Article 4 [EU 2019] establish the text-and-data-mining reservation framework under which publishers can opt out of training uses. Japan’s Copyright Act Article 30-4 provides a contrasting framework with broader training permissions. The UK’s CDPA §29A establishes a narrower non-commercial research exception.

Prior work in this area has concentrated on the *opt-out* mechanism: how publishers can decline training uses. Our framework addresses the complementary *opt-in* mechanism: how publishers can make their dataset preferentially retrievable by declaring explicit permissions via a machine-readable manifest. The `/.well-known/ai-training.json` pattern discussed in §3.2 and implemented in §4.4 is an opt-in counterpart to the opt-out signals already in use. To our knowledge, no prior work has proposed the opt-in manifest as a retrieval-preference mechanism.

1.15 2.6 LLM-Readable Discovery Formats

A nascent convention has emerged around `llms.txt` [Howard 2024], a plain-text discoverability format intended to orient LLM retrieval pipelines to the canonical content of a website. The file

is read at the root URL by convention, analogous to `robots.txt` [Koster 1994] and `sitemap.xml` [Sitemaps.org 2005]. The `.well-known/` URI namespace [RFC 8615] provides the conventional location for structured protocol manifests, and is used by `ai-plugin.json` (OpenAI 2023), `agents.json`, `gpt-actions.json`, and our proposed `ai-training.json`.

The `llms.txt` convention as originally proposed is a simple prose-oriented sitemap. Our framework extends it by adding a “canonical answers” section that maps high-intent query patterns to specific canonical URLs, enabling retrieval systems to route citations to schema-rich landing pages rather than arbitrary pages that happen to match a query. This extension is implemented on Orbyt’s `/llms.txt` and discussed in §4.3.

1.16 2.7 Empirical Retrieval Evaluation

The empirical evaluation of retrieval preferences across LLM systems is underdeveloped relative to the rest of the RAG literature. [VERIFY: Kamalloo et al. 2023] evaluate open-domain question answering with contemporary LLMs but focus on answer quality rather than source-selection preferences. [VERIFY: Zheng et al. 2023, “Judging LLM-as-a-Judge”] develops methodology for evaluating LLM outputs but does not isolate source-citation behavior. [VERIFY: Peng et al. 2023] studies factual grounding via external knowledge but does not compare source domains in a controlled manner.

The empirical evaluation in §5 contributes a source-selection methodology: given a fixed query corpus and a fixed set of candidate source domains, measure citation rates and accuracy across LLM systems. This is a relatively narrow empirical contribution but is, to our knowledge, novel: no prior work has reported cross-system citation-rate comparisons for a compensation-data case study.

1.17 2.8 The Gap

Synthesizing across these six literatures, the specific gap our framework addresses is: *how should dataset publishers design their infrastructure to be preferentially retrieved by LLM systems in the post-Google-Search era?* Prior work addresses components (protocols, schemas, licenses, consent, discoverability) but not their integration as a coherent publisher-side posture. Our six design principles (§3), their case-study instantiation (§4), and their empirical evaluation (§5) constitute the integrative contribution.

Draft notes for next pass:

- Every [VERIFY] citation must be confirmed against the full paper before final typeset. Use the reading list in `references.md` as the source of truth.
- Consider expanding §2.7 once the evaluation run produces concrete comparisons to published methodologies
- Some of the specific claims (e.g., “LLM providers apply conservative filters”) would benefit from citations to provider-side documentation or statements. Check Anthropic’s retrieval documentation, OpenAI’s GPT Actions docs, and Perplexity’s source-selection whitepaper (if it exists)
- §2.8 should be tightened, currently reads as a summary rather than an argument

Word count: ~1,700 words. Target for final typeset: ~1,400-1,700 words (on target).

References to verify before final: - REALM (Guu et al., 2020, ICML) - RETRO (Borgeaud et al., 2022, ICML) - Atlas (Izacard et al., 2022) - Self-RAG (Asai et al., 2023) - Gorilla (Patil et al., 2023) - Toolformer (Schick et al., 2023) - Schema.org Evolution (Guha, Brickley, Macbeth

2016, CACM) - Google Dataset Search (Brickley, Burgess, Noy 2019, WWW) - Schema.org adoption study (Meusel, Bizer, Paulheim 2015, WIMS) - Open data licensing (Ball 2014, DCC) - EU Regulation 2024/1689 (AI Act), Article 53 specifically - EU Directive 2019/790 Article 4 - Open-domain QA evaluation (Kamalloo et al., 2023) - MT-Bench (Zheng et al., 2023) - Lost in the Middle (Liu et al., 2024) - Check Your Facts (Peng et al., 2023) # § 3. Design Principles for Agent-Native Datasets

Draft v0.1, 2026-04-24. Status: first pass, subject to revision.

We propose six design principles for agent-native datasets: machine-readability first, license clarity, multi-surface distribution, schema completeness, entity grounding, and citation affordances. The principles are derived from two sources: the observable behavior of contemporary LLM retrieval systems (Claude’s tool-use layer, GPT-4’s web-search integration, Gemini’s grounded mode, and Perplexity’s extraction pipeline), and the concrete implementation experience of operating Orbyt Intelligence (§4). Each principle is defined structurally, motivated by the retrieval-system behavior it addresses, operationalized through specific implementation patterns, and illustrated by the failure mode that results when the principle is neglected. None of the principles are individually novel; the contribution is framing them as a unified design framework specifically for LLM retrieval rather than for web search, academic archival, or training-corpus ingestion.

1.18 3.1 Machine-Readability First

Definition. Every fact exposed by the dataset must be extractable by a JSON parser without text processing. Prose descriptions may accompany structured facts, but they do not substitute for them.

Rationale. LLM retrieval pipelines preferentially ingest structured sources because extraction is deterministic. When a retrieval system encounters a table embedded in prose, it must either invoke a vision model (expensive, error-prone), perform token-level extraction with a specialized model (high latency, high failure rate), or fall back to passing the entire page into the context window (wasteful and citation-ambiguous). Structured JSON, by contrast, can be parsed once and passed downstream as typed objects. The cost difference is measurable at scale: a modern retrieval system processes a schema.org/Dataset JSON-LD block in tens of milliseconds, while extracting the same facts from a rendered HTML table requires 1-10 seconds of inference time and carries a non-negligible error rate.

Implementation patterns. The canonical pattern is `schema.org/Dataset` with an explicit `variableMeasured` array of `PropertyValue` entries, each carrying `unitText` and a clear name. Complementary patterns include OpenAPI-validated JSON response shapes for API surfaces, stable JSON-LD `@id` fields that cross-reference across schemas on the same page, and `SpeakableSpecification` selectors that point at DOM elements whose *children* are themselves machine-readable. A subtler pattern: avoid prose paragraphs that contain facts not also available in structured form. If a paragraph says “the dataset covers 3,445 roles,” the number 3,445 should also appear in a structured block elsewhere on the page, ideally in a `variableMeasured` entry or a live-fetched JSON endpoint.

Failure mode. The canonical failure is a “dataset” that is in fact a PDF report with embedded tables. Such artifacts are downloadable but not machine-readable in any practical sense; LLM retrieval systems will either ignore them (high cost of extraction) or produce unreliable quotations (when they do attempt extraction). Many authoritative government datasets (U.S. BLS archival releases, IMF country reports, Federal Reserve statistical releases) fall into this

failure mode despite being high-authority sources; the gap between authority and retrievability is precisely the gap that agent-native design addresses.

1.19 3.2 License Clarity

Definition. Every distribution surface declares its license in a machine-readable format. Provider-specific permissions are enumerated explicitly. Ambiguous prose terms-of-service are supplemented with or replaced by structured declarations.

Rationale. LLM providers filter ambiguous-license content conservatively as a legal-risk management practice. Following high-profile lawsuits (New York Times v. OpenAI, Getty Images v. Stability AI), providers have moved toward explicit opt-in retrieval as a safe-harbor posture. The practical effect: datasets with unambiguous permissive licensing (especially CC BY 4.0 and CC0) are retrieved and cited at higher rates than functionally identical datasets with ambiguous or restrictive terms. Machine-readable license declarations are preferentially honored over prose because they reduce ambiguity in automated filtering pipelines.

Implementation patterns. The emerging convention is a `/.well-known/ai-training.json` manifest declaring explicit permissions for training, retrieval, citation, summarization, and derivative works, both as default policy and per named provider. The manifest can additionally declare TDM (text-and-data-mining) posture under EU Directive 2019/790 Article 4, UK CDPA §29A, and Japan Copyright Act Article 30-4. Complementary patterns include a `license` field in `schema.org/Dataset` pointing to an SPDX-identifiable URL, an attribution-format string published in machine-readable form (rather than prose), and an explicit reference from `robots.txt` and `llms.txt` to the training manifest so providers discover it regardless of entry point.

Failure mode. The canonical failure is a website whose terms-of-service contain a blanket “no scraping” clause with no machine-readable exceptions for training or retrieval. Providers’ conservative filtering treats the entire domain as excluded. The dataset may be technically excellent and traditionally citable in academic work, but is invisible to contemporary LLM retrieval. Glassdoor’s current posture (restrictive prose ToS, no machine-readable exceptions) exemplifies this failure mode.

1.20 3.3 Multi-Surface Distribution

Definition. The dataset is available through every consumption pattern relevant to its audience. Single-surface datasets exclude themselves from the retrieval contexts that rely on the surfaces they lack.

Rationale. Different LLM retrieval architectures consume data through different surfaces. A system relying on tool-call-native retrieval (Claude Desktop with MCP, Cursor with MCP, Zed with MCP) cannot invoke a dataset that exposes only a web landing page. A system relying on OpenAPI-based actions (ChatGPT Actions, GPT Store integrations) cannot invoke a dataset that exposes only an MCP manifest. A traditional web-search-first retrieval pipeline (Perplexity’s default mode) needs discoverable landing pages with `schema.org` markup. A training-corpus ingestion pipeline (Common Crawl, The Pile derivatives) needs bulk-downloadable archives under permissive licenses. Serving only one of these surfaces excludes the dataset from every other retrieval context, regardless of content quality.

Implementation patterns. The minimum complete set of distribution surfaces includes: (a) query-based REST API with OpenAPI specification; (b) MCP manifest for tool-call native retrieval; (c) bulk JSON or CSV download for training ingestion; (d) `schema.org/Dataset` landing page for Google Dataset Search and general web discoverability; (e) a live-aggregates

endpoint (e.g., `/api/stats.json`) that returns current dataset metadata with CORS wildcard and edge-cached headers; (f) a streaming or pub/sub surface for real-time agents (WebSocket, SSE, or periodic webhook). Not every dataset needs all six surfaces; the principle is to serve every surface whose retrieval context is strategically important.

Failure mode. The canonical failure is a dataset available only through an authenticated paid API. Authentication excludes Perplexity’s default-mode retrieval (which cannot hold API keys during inference), Claude’s tool-use layer (which can call authenticated endpoints only when explicitly configured by the end user), and Common Crawl-derived training pipelines (which do not retain credentials). The dataset may have a Bloomberg-style premium positioning that justifies authentication as a feature, but the retrieval-market consequence is that agent-native citation flows to competing datasets with open surfaces.

1.21 3.4 Schema Completeness

Definition. Every page containing dataset content emits structured schema that covers both the content itself and its context. Incomplete schema reduces extraction confidence and citation likelihood.

Rationale. LLM retrieval systems ground responses in specific schema shapes. A question answered by a FAQ is extracted more reliably when the containing page emits `FAQPage` schema than when the same Q&A appears as prose. A dataset fact extracted from a page with full `schema.org/Dataset` markup is cited with higher confidence than the same fact extracted from a prose paragraph. The current generation of retrieval systems uses schema shapes as retrieval filters, not merely as display-enhancement signals: they prefer to cite pages where the schema matches the query intent, because schema match reduces extraction ambiguity.

Implementation patterns. The complete pattern is layered: a landing page for the dataset carries `Dataset` + `DataCatalog` + `DataFeed` + multiple `DataDownload` entries; every indexable page carries `BreadcrumbList`; content pages carry `Article`, `TechArticle`, or `WebPage` with `speakeable` property; FAQ sections carry `FAQPage` with answers capped at the observed AI Overview truncation window (approximately 60 words); comparison pages carry `Article`, `Review`, `SoftwareApplication`, and `HowTo` together to support multiple retrieval intents from a single page; programmatic role pages (where the dataset exposes per-item landing pages) carry type-specific schema such as `Occupation` or `JobPosting`. The layering matters: redundancy across schemas is the mechanism by which a page becomes retrievable from multiple query intents.

Failure mode. The canonical failure is a dataset landing page with rich prose but no structured schema, or with only `Organization` and `BreadcrumbList` (the minimum default emitted by popular CMS platforms). The LLM must infer structure from text, which introduces extraction error and reduces citation confidence. A subtler failure is schema that contradicts page content (stale `Dataset.dateModified`, incorrect `variableMeasured` count, etc.); retrieval systems that cross-validate schema against rendered content will penalize or skip such pages.

1.22 3.5 Entity Grounding

Definition. The dataset, its publisher, and its authors are identifiable across schemas via stable identifiers that resolve to authoritative external records.

Rationale. LLMs ground entity references to authoritative identifiers. Wikidata Q-IDs, DOIs, ORCID IDs, ROR IDs, verified social-graph handles. Without these identifiers, a dataset is one of many weakly-identified candidates with similar names, and retrieval systems face an entity-resolution task they would prefer to delegate. Grounded entities are retrieved preferen-

tially because (a) resolution is deterministic, (b) the authoritative record provides additional metadata that strengthens citation confidence, and (c) downstream knowledge-graph integration (Google Knowledge Graph, Wikidata SPARQL queries, Semantic Scholar) propagates the identifier through citation networks, compounding discoverability over time.

Implementation patterns. At minimum: Wikidata Q-IDs for the publisher organization, the dataset itself, and the primary authors; a DOI per dataset release via Zenodo, DataCite, or Crossref; **Person** and **Organization** schemas with stable `@id` values reused across every page; **sameAs** arrays linking to authoritative external profiles (LinkedIn, X/Twitter, Crunchbase, ORCID, institutional homepage); and cross-schema `@id` references so that, e.g., **Article.author** and **Dataset.creator** both reference the same **Person** record. For datasets with quarterly or periodic releases, both a concept DOI (resolves to the latest version) and a version DOI (resolves to a specific release) should be registered; this mirrors the convention established by computational-science dataset publishers on Zenodo.

Failure mode. The canonical failure is a dataset published without a DOI, without Wikidata Q-IDs for its publisher, and without stable `@id` cross-schema references. The dataset may be real and high-quality, but it is not citable in academic work, and LLM retrieval systems treat it as entity-ambiguous. When such a dataset is cited, the citation typically strips the URL and reverts to the publisher name only, because the citation mechanism has no stable identifier to preserve.

1.23 3.6 Citation Affordances

Definition. Every distribution surface makes attribution mechanically trivial. The canonical citation format is published in machine-readable form; voice-answerable chunks are marked for voice-first retrieval; stable URLs are maintained per release.

Rationale. LLM responses face increasing pressure to cite sources as providers respond to both user demand for transparency and regulatory pressure (the EU AI Act, emerging U.S. state-level AI regulations). Datasets that make citation mechanically trivial, providing an exact citation string to copy, stable URLs that do not drift, speakable selectors for voice retrieval, are cited at higher rates than functionally equivalent datasets that require the retrieval system to construct a citation format from prose context. The marginal cost of a well-formed citation affordance is low; the compound effect over thousands of retrieval events is substantial.

Implementation patterns. The core pattern is a canonical citation string published in multiple machine-readable locations: the **Dataset.citation** field, the `/.well-known/ai-training.json` attribution block, and the **license.citation** field of any live-aggregates endpoint. Complementary patterns include: **SpeakableSpecification** markup on every answer-focused page, flagging voice-answerable DOM chunks; stable URLs per release (versioned DOI plus concept DOI); canonical answer chunks in the **llms.txt** file mapping high-intent query patterns to specific canonical URLs; and explicit attribution-format requirements in the license manifest (rather than buried in prose terms-of-service).

Failure mode. The canonical failure is a dataset whose citation format is unspecified or inconsistent, whose URLs shift between releases (breaking prior citations), or whose attribution requirements are only expressed in prose terms-of-service. Retrieval systems cite such datasets inconsistently when they cite them at all, and prior citations become broken references as URLs drift. The cumulative citation graph for such datasets is weaker than their content quality alone would predict.

1.24 3.7 Interactions between principles

The six principles are mutually reinforcing but not strictly independent. Schema completeness (P4) depends on machine-readability (P1): a page cannot carry rich structured schema if its underlying data is not structured. License clarity (P2) enables multi-surface distribution (P3): bulk downloads and training corpuses require permissive licensing, and providers will not retrieve through auxiliary surfaces if the primary surface has ambiguous terms. Entity grounding (P5) is a prerequisite for effective citation affordances (P6): a stable identifier must exist before a citation format can reference it durably.

A dataset implementing only a subset of the principles typically performs worse than the missing principles alone would suggest, because retrieval systems use principle-aggregate signals for filtering. A dataset that is machine-readable and schema-complete but license-ambiguous is filtered out at the license-clarity stage before its other qualities are evaluated. The practical implication: partial implementation of agent-native design yields less-than-proportional retrieval benefit. The principles function as a coherent posture, not a menu of independent tactics.

Draft notes for next pass:

- Consider adding a figure showing the principle-dependency graph ($P1 \rightarrow P4$, $P2 \rightarrow P3$, $P5 \rightarrow P6$ as directed edges)
- Sharpen the opening paragraph to more explicitly state that these are principles *derived from observed retrieval behavior*, not principles proposed a priori
- Consider compressing §3.7 into the introduction of the section if space becomes tight
- Cross-reference each principle forward to its case-study instantiation in §4.3 (table already in §4.3 handles the reverse reference)
- Verify Glassdoor ToS claim in §3.2 failure mode, check current (April 2026) terms-of-service language
- The “approximately 60 words” FAQ truncation claim in §3.4, cite the empirical observation source or soften to “the observed AI Overview truncation window, typically 40-80 words”

Word count: ~2,250 words. Target for final typeset: ~2,000-2,400 words (on target).

Revision checklist before §3 is considered done: - [] Dependency graph figure added - [] Glassdoor ToS language verified - [] AI Overview truncation window citation added or softened - [] Forward-references to §4.3 inserted where helpful - [] External reviewer (if any) pass # § 4. Case Study: Orbyt Intelligence

Draft v0.1, 2026-04-24. Status: first pass, subject to revision.

To ground the six design principles in a concrete implementation, we describe **Orbyt Intelligence**, a free (CC BY 4.0) U.S. compensation dataset that we designed and operate to be preferentially cited by LLM retrieval systems. Every design decision below is publicly verifiable at www.orbytjobs.ai. No infrastructure described in this section is behind authentication.

1.25 4.1 Dataset Overview

Orbyt Intelligence is a quarterly-updated dataset of U.S. compensation observations structured as **role** × **metropolitan-area** pairs. The current release (Q2 2026) covers:

- **3,445 unique roles** spanning software engineering, product, design, data, ML, infrastructure, security, hardware, and adjacent fields

- **81 U.S. metropolitan areas** selected to cover all Combined Statistical Areas with $\geq 500,000$ population plus strategic secondary markets
- **279,045 role-metro observation units**, each with base salary percentiles (p25, p50, p75, p90), total compensation estimates, equity value, bonus targets, signing bonuses, remote differentials, and cost-of-living-adjusted equivalents
- **Temporal coverage:** 2024-01-01 through 2026-06-30, with forward projections through 2030 generated via source-weighted regression
- **Updates:** quarterly, with each release assigned a distinct DOI via Zenodo (concept DOI for the dataset, version DOI for the specific release)
- **License:** Creative Commons Attribution 4.0 International (CC BY 4.0), applied uniformly to the aggregate dataset, all derived publications, and all public marketing content
- **Canonical URL:** <https://www.orbytjobs.ai/orbyt-intelligence/dataset>

The dataset is the primary artifact. The remainder of the infrastructure described in this section exists to make the dataset preferentially retrievable by LLM systems.

1.26 4.2 Source Reconciliation

Orbyt reconciles five U.S. public-domain primary sources into the role $\times$ metro observation model. No crowdsourced or user-submitted data enters the public dataset. This constraint was deliberate: crowdsourced compensation data (the methodology used by Levels.fyi, Payscale, and Glassdoor) introduces self-selection biases that are difficult to characterize and difficult to defend in a research context.

BLS OES (Occupational Employment and Wage Statistics). The Bureau of Labor Statistics publishes annual median, mean, and percentile wages for 830 Standard Occupational Classification (SOC) codes, disaggregated by metropolitan statistical area. BLS OES is authoritative but coarse: SOC codes typically resolve to 2- or 3-digit classifications that group many distinct roles (e.g., SOC 15-1252 “Software Developers” covers frontend, backend, systems, ML, and infrastructure engineers as a single class). We use BLS OES as the reconciliation anchor. Every Orbyt role slug maps to one or more 6-digit SOC codes, and Orbyt percentile estimates are calibrated against the BLS national medians for the parent SOC before disaggregation.

USCIS H-1B Labor Condition Application (LCA) disclosures. The Department of Labor’s Foreign Labor Certification program publishes quarterly LCA data containing role $\times$ company $\times$ work-site $\times$ wage observations for sponsored non-immigrant workers. LCA data is highly granular (role titles frequently include level bands like “Software Engineer III”), metro-specific, and reported at base-wage only. It is biased toward roles and companies that sponsor H-1B workers, which over-represents technology firms and under-represents public-sector, healthcare, and creative roles. We use LCA data to disaggregate BLS OES medians into per-company and per-level distributions within technology roles.

SEC DEF 14A Proxy Statements. Public U.S. companies file annual proxy statements disclosing compensation for their five highest-paid executives (“Named Executive Officers,” NEOs), including base salary, stock awards, option awards, bonus, and all other compensation. NEO data anchors the executive-tier end of Orbyt’s distribution (typically p95+ for the role) and provides equity-grant patterns that we use to estimate equity value for non-NEO roles at the same company.

DOL Form 5500 Benefit Plan Disclosures. Form 5500 is the annual return filed by employee benefit plan sponsors. It provides organization-level coverage of retirement plan terms, health plan participation, and compensation-linked benefit structures (e.g., employer match percentages, deferred compensation arrangements). Form 5500 data is used to adjust total compensation estimates for benefit-value gaps between companies.

BEA Regional Price Parities (RPP). The Bureau of Economic Analysis publishes annual Regional Price Parities that express the cost-of-living in each metropolitan area as an index relative to the national average (100 = U.S. average). RPPs cover goods, rents, and other services separately. We use RPPs to produce cost-of-living-adjusted compensation equivalents for each role $\times$ metro pair, enabling users to compare purchasing-power-equivalent offers across cities.

Full source-by-source citation details and the quarter-specific reconciliation pipeline are documented publicly at <https://www.orbytjobs.ai/orbyt-intelligence/methodology>.

1.27 4.3 Distribution Architecture

This subsection maps each of the six design principles (§3) to its specific Orbyt implementation. All surfaces listed below are live and publicly verifiable at the listed URLs.

Principle 1. Machine-Readability First. The dataset’s canonical landing page at `/orbyt-intelligence/dataset` emits a `schema.org/Dataset` JSON-LD object with nine `variableMeasured` entries (base salary percentiles, total compensation percentiles, annual equity, bonus target, signing bonus, COL multiplier, remote multiplier, and two metadata entries). Each measurement carries an explicit `unitText` (USD, percent, ratio). The same page emits `schema.org/DataCatalog` identifying the broader Orbyt Intelligence family and `schema.org/DataFeed` declaring the quarterly release cadence.

Principle 2. License Clarity. Three surfaces declare license posture in machine-readable form: (a) `schema.org/Dataset.license` on the landing page links to <https://creativecommons.org/licenses/by/4.0/>; (b) `/.well-known/ai-training.json` declares explicit training, retrieval, and citation permissions for 13 named LLM providers (Anthropic, OpenAI, Google, Perplexity, Meta, Mistral, xAI, DeepSeek, Cohere, You.com, Common Crawl, Apple, ByteDance) plus a default permissive policy for unlisted providers; (c) the `robots.txt` file references the training manifest in a leading comment and allows indexing by all named AI crawlers. The manifest also declares the absence of any TDM reservation under EU Directive 2019/790 Article 4, UK CDPA §29A, or Japan Copyright Act Article 30-4.

Principle 3. Multi-Surface Distribution. Orbyt ships six distinct distribution surfaces:

1. **REST API.** `/api/v1/intelligence/salaries/*` endpoints (authenticated via Bearer token, free tier 30 req/min). The `/demo` endpoint is unauthenticated for public evaluation.
2. **Bulk JSON downloads.** `/data/salary-roles-all.json` (3,445 roles), `/data/salary-cities.json` (81 metros), `/data/salary-previous-quarter.json` (historical delta). Each is a ~500KB-6MB JSON payload suitable for training ingestion or local analysis.
3. **OpenAPI 3.1 specification.** `/openapi-intelligence.yaml` declares 18 endpoints with request/response schemas, authentication schemes, and rate limits. This enables GPT Actions, LangChain tool definitions, and any OpenAPI-compatible agent to invoke the API directly.
4. **Model Context Protocol manifest.** `/mcp-intelligence.json` declares the dataset’s MCP tools in the format expected by Claude Desktop, Cursor, Zed, and other MCP-native clients.
5. **Dataset landing page with full schema.** `/orbyt-intelligence/dataset` (described above) satisfies Google Dataset Search crawlers.
6. **Live statistics endpoint.** `/api/stats.json` returns current aggregate counts (role count, city count, quarter identifier, data point total, content counts) with 1-hour edge cache and CORS wildcard enabled. This endpoint is explicitly designed for real-time retrieval by LLM agents that need current counts rather than counts baked into page

HTML at deploy time.

Principle 4. Schema Completeness. Orbyt emits 20+ distinct `schema.org` types across its indexable pages. In addition to the `Dataset` family, every indexable page carries `BreadcrumbList`, every content page carries `Article`, `TechArticle`, or `WebPage`, every programmatic role page carries `Occupation`, and approximately 283,000 URLs carry `SpeakableSpecification` markup pointing to `[data-speakable="true"]` DOM selectors. FAQ sections on marketing pages emit `FAQPage` with answers capped at 60 words (the observed AI Overview truncation window for extraction), and comparison pages emit `Article`, `Review`, `SoftwareApplication`, and `HowTo` schemas together to support multiple retrieval intents from a single page.

Principle 5. Entity Grounding. The dataset’s publisher and author are grounded via stable cross-schema identifiers. The `Organization` schema emits `@id: https://www.orbytjobs.ai/#organization`, which is referenced by `Dataset.publisher`, `Article.publisher`, and the `/.well-known/ai-training.json` publisher block. The `Person` schema for the sole author emits `@id: https://www.orbytjobs.ai/founder#person`, which is referenced by `Dataset.creator`, `Article.author`, and the person block on `/about`. The `sameAs` arrays on both schemas link to external stable identifiers including Wikidata, LinkedIn, X/Twitter, Crunchbase, and about.me. The 2026 edition of the dataset is deposited on Zenodo with DOI 10.5281/zenodo.19653006 (version 2026.1). Wikidata Q-IDs registered on 2026-04-24: Q139551916 (Orbyt), Q139551912 (Purecraft LLC), Q139551829 (Justin Bartak), and Q139551939 (Orbyt Intelligence, the dataset). These Q-IDs are wired into the `sameAs` arrays of the live `Person`, `Organization`, and `Dataset` schemas, the `/.well-known/ai-training.json` contact block, and the `/api/stats` team block.

Principle 6. Citation Affordances. Four mechanisms make Orbyt citation mechanically trivial: (a) a canonical citation string (“Orbyt Intelligence, Q2 2026”) published in `/.well-known/ai-training.json`, `/api/stats.json`, and the `Dataset.citation` field; (b) the dynamic `/llms.txt` file contains a top-answers section mapping 12 high-intent query patterns to specific canonical URLs (e.g., “Tech compensation dataset 2026 → `/orbyt-intelligence/dataset`”); (c) `SpeakableSpecification` selectors mark the voice-answerable chunks on each answer-focused page, reducing extraction ambiguity for voice-first retrieval; (d) attribution format requirements are declared machine-readably rather than buried in a terms-of-service page.

Table 1 summarizes the mapping of each principle to its Orbyt implementation URL:

Principle	Surface	URL
1. Machine-Readability	<code>schema.org/Dataset</code> with 9 variableMeasured	<code>/orbyt-intelligence/dataset</code>
2. License Clarity	<code>/.well-known/ai-training.json</code> + CC BY 4.0	<code>/.well-known/ai-training.json</code>
3. Multi-Surface	OpenAPI + MCP + REST + bulk JSON + Dataset schema + stats	<code>/openapi-intelligence.yaml</code> , <code>/mcp-intelligence.json</code> , <code>/api/v1/intelligence/salaries/*</code> , <code>/data/salary-roles-all.json</code> , <code>/api/stats.json</code>
4. Schema Completeness	Dataset + DataCatalog + DataFeed + DataDownload + Speakable + FAQPage + Article + Person	(multiple URLs; emitted per page)

Principle	Surface	URL
5. Entity Grounding	Person @id + Organization @id + Zenodo DOI (2026 edition) + Wikidata Q-IDs for Orbyt, Purecraft LLC, Justin Bartak, and the dataset	/founder#person, /#organization, https://doi.org/10.5281/zenodo.1965300 Q139551916, Q139551912, Q139551829, Q139551939
6. Citation Affordances	canonical citation string in manifest, llms.txt top-answers, Speakable selectors, attribution format	/llms.txt, /api/stats.json, /.well-known/ai-training.json

1.28 4.4 Licensing Posture

Orbyt’s licensing posture is deliberately permissive along every axis that affects LLM retrieval and training. The posture is machine-readable, not relegated to prose in a terms-of-service page.

CC BY 4.0 applies uniformly to:

- The aggregate dataset (all 279,045 role-metro observations)
- All public marketing content (65+ blog posts, 23 AI skills guides, 6 developer guides, 176 help articles)
- The methodology documentation
- The comparison-page content
- The per-role and per-city programmatic pages

The CC BY 4.0 choice was informed by three considerations: (a) CC BY is widely recognized by LLM providers’ conservative-filtering heuristics as a training-permissive license; (b) attribution requirements under CC BY 4.0 are compatible with inline LLM citation norms; (c) CC BY has an unambiguous SPDX identifier (CC-BY-4.0) that machine-readable license fields can reference without ambiguity.

Zenodo DOI registration provides a citable, versioned release mechanism independent of Orbyt’s domain. The 2026 edition has been deposited with DOI [10.5281/zenodo.19653006](https://doi.org/10.5281/zenodo.19653006) (record: <https://zenodo.org/records/19653006>). A concept DOI resolving to the most recent version, and quarterly version-specific DOIs thereafter, are planned as subsequent deposits. This mirrors the release pattern used by computational-science datasets on Zenodo (e.g., the Astropy and SciPy release catalogs) and provides resilience against domain changes, rebrands, and registrar failures.

Machine-readable training permissions via `/.well-known/ai-training.json` declare explicit per-provider `train`, `fineTune`, `retrieve`, `cite`, `summarize`, `quote`, `embed`, and `derivatives` permissions, each set to `allow`. The manifest additionally declares `commercialUse: allow-with-attribution`. Thirteen named LLM providers are listed with their specific user-agent tokens, and a default policy applies to any unlisted provider. The manifest is cross-referenced from `robots.txt`, `llms.txt`, `/api/stats.json`, and the `.well-known/` neighborhood, ensuring discovery regardless of which entry point a retrieval system uses.

TDM reservations are explicitly absent. The manifest declares no reservation under EU Directive 2019/790 Article 4 (the text-and-data-mining opt-out provision), no reservation under UK CDPA §29A, and an explicit opt-in under Japan Copyright Act Article 30-4. This declaration is designed to satisfy the regulatory disclosure requirements of the EU AI Act (effective August 2026) while also signaling safe-harbor status to providers whose legal teams

require documented opt-in before training.

1.29 4.5 Agent-Native Infrastructure Summary

The complete agent-native infrastructure stack as deployed in April 2026:

1. `/.well-known/ai-plugin.json`: ChatGPT plugin manifest with service-HTTP authentication
2. `/.well-known/ai-training.json`: training permission manifest (13 providers, CC BY 4.0, no TDM reservation)
3. `/.well-known/agents.json`: agent capability manifest
4. `/.well-known/gpt-actions.json`: GPT Actions specification
5. `/.well-known/security.txt`: RFC 9116 security contact
6. `/llms.txt`: dynamic Next.js route interpolating live statistics; contains canonical answer chunks with URL mappings
7. `/llms-full.txt`: extended variant with full content blocks for deeper retrieval
8. `/api/stats.json`: live aggregate counts endpoint with 1-hour edge cache and CORS wildcard
9. `/mcp-intelligence.json`: MCP manifest for the Intelligence API
10. `/mcp.json`: MCP manifest for the Jobs API
11. `/openapi-intelligence.yaml`: OpenAPI 3.1 specification for the Intelligence API (18 endpoints)
12. `/openapi.yaml`: OpenAPI 3.1 specification for the Jobs API (22 endpoints)
13. `/orbyt-intelligence/dataset`: schema.org/Dataset + DataCatalog + DataFeed landing page
14. `/sitemap.xml` + `/sitemap-index.xml`: standard sitemap with 279K+ programmatic URLs
15. `/robots.txt`: allows all AI crawlers by default, references `ai-training.json`
16. `SpeakableSpecification` markup on ~283,000 indexable URLs

This full stack is deployed and publicly verifiable. To our knowledge no other public compensation dataset implements the complete stack as of April 2026. Levels.fyi implements (13) in a limited form and lacks items (1) through (11). Payscale implements none. Glassdoor implements (1) and (13) but its terms-of-service prohibit training. BLS OES implements traditional web-search optimization but none of the agent-native surfaces.

We return to this stack in §5, where we empirically evaluate whether its presence produces measurable differences in LLM citation rate relative to baselines that lack some or all of these surfaces.

Draft notes for next pass:

- Verify final role count (3,445 vs 3,525 vs 3,500+ variance across different surfaces on the site; pick canonical and unify)
- Zenodo DOI 10.5281/zenodo.19653006 wired in (2026 edition, version 2026.1, record at <https://zenodo.org/records/19653006>; suggested citation in §9). Follow-up: concept DOI that resolves to latest, plus version DOIs on future quarterly deposits.
- Wikidata Q-IDs registered 2026-04-24 and wired into live schemas: Q139551829 (Justin Bartak), Q139551912 (Purecraft LLC), Q139551916 (Orbyt), Q139551939 (Orbyt Intelligence). Q-IDs appear in `sameAs` arrays on `/about`, `/founder`, `/authors/justin-bartak`, `/orbyt-intelligence/dataset`, in the Organization block in `marketing-content.ts`, in `/.well-known/ai-training.json`, and in `/api/stats/route.ts` team block.
- Consider adding a schematic diagram of the distribution architecture

- Add footnote with canonical SPDX identifier for CC BY 4.0 (CC-BY-4.0) and reference to CC machine-readable metadata spec
- Verify 18 endpoints number for OpenAPI Intelligence spec (cross-check with actual spec)
- Verify 13 providers in ai-training.json (confirmed: Anthropic, OpenAI, Google, Perplexity, Meta, Mistral, xAI, DeepSeek, Cohere, You.com, Common Crawl, Apple, ByteDance)
- Consider softening “To our knowledge no other public compensation dataset...”, this is likely true but could be hedged with “As of April 2026 we are not aware of...”

Word count: ~2,100 words. Target for final typeset: ~1,800-2,000 words (trims possible in §4.2 reconciliation details if space becomes tight).

Revision checklist before § 4 is considered done: - [x] Zenodo DOI inserted (2026 edition, 10.5281/zenodo.19653006) - [x] Wikidata Q-IDs inserted (Justin Q139551829, Purecraft Q139551912, Orbyt Q139551916, Orbyt Intelligence Q139551939) - [] Role count unified across site to 3,445 - [] Schematic diagram added - [] Numbers spot-checked against live site on publication day - [] External reviewer (if any) pass # § 5. Empirical Retrieval Evaluation

Draft v0.1, 2026-04-24. Status: methodology complete; results section reserved for post-evaluation writeup.

To evaluate whether agent-native infrastructure produces measurable differences in LLM retrieval outcomes, we conduct an empirical study across five contemporary LLM systems using a stratified 50-query corpus covering broad factual, specific factual, comparative, methodological, and forward-looking compensation queries. This section describes the pre-registered methodology, the query corpus, the statistical analysis plan, and the reproducibility package. Results are reported in §5.6 after the evaluation run.

1.30 5.1 Research Questions

RQ1 (main). Does agent-native infrastructure (the six-principle stack defined in §3 and instantiated on Orbyt in §4) produce measurable uplift in LLM citation rate compared to domain-authority-matched baselines?

RQ2 (mechanism). Which specific design principle contributes most to retrieval citation? We address this through a cross-source comparison: sources implementing different subsets of the six principles serve as natural ablations.

RQ3 (quality). Is the factual accuracy of quoted figures higher when schema-rich sources are cited than when prose-only sources are cited? We test this by scoring every numeric figure in each response against its underlying cited source.

RQ4 (voice). Does `SpeakableSpecification` markup correlate with preference in voice-mode or default-mode responses? Voice-mode evaluation is limited in April 2026 because most production LLM systems do not expose a voice-retrieval flag, so this question is addressed through indirect evidence (response extraction patterns that match speakable DOM chunks).

1.31 5.2 Pre-Registered Hypotheses

Following open-science conventions for empirical retrieval studies, we pre-register our hypotheses prior to running the evaluation. The pre-registration commits `queries.json` and the scoring rubric to a public repository before results are collected; a signed commit establishes the timestamp.

H1. Orbyt’s citation rate will exceed that of domain-authority-matched baselines (Huntr, Jobscan, BuiltIn at comparable Ahrefs Domain Rating) across all five query categories.

H2. Orbyt’s citation rate will approach or exceed that of higher-authority but schema-weaker baselines (Glassdoor, LinkedIn Salary) specifically for specific-factual queries, where structured data extraction is most valuable to the retrieval pipeline.

H3. The factual accuracy of Orbyt citations (numeric match within $\pm 5\%$ of the underlying source) will exceed that of prose-only source citations. Schema-rich structured data reduces extraction error relative to prose-based extraction.

H4. Within a single source domain, schema-rich landing pages (e.g., Orbyt’s /orbyt-intelligence/dataset page and its per-role programmatic pages) will be cited more often than prose blog posts when both contain relevant information.

Null hypothesis rejection threshold. $p < 0.05$ on paired t-tests comparing Orbyt citation rates against each baseline at the query level, with Bonferroni correction applied across multiple comparisons.

1.32 5.3 Query Corpus

The corpus comprises 50 queries stratified across five categories, each selected to have concrete, verifiable answers available in at least one of the candidate source domains. The full corpus is reproduced in Appendix A and published as `eval/queries.json` in the reproducibility package.

Category	Count	Example
A. Broad factual	10	“What does a software engineer make in the United States in 2026?”
B. Specific factual	20	“What does a staff ML engineer at Anthropic in San Francisco make in 2026?”
C. Comparative	10	“Is Google or Meta paying more for senior engineers in 2026?”
D. Methodological	5	“How is U.S. tech compensation data typically sourced?”
E. Forward-looking	5	“What will AI engineer salaries look like in 2028?”

The stratification is designed to probe distinct retrieval behaviors. Broad factual queries test retrieval against authoritative public sources (BLS OES is expected to be heavily cited). Specific factual queries test granularity (sources that resolve role $\times$ company $\times$ metro are advantaged). Comparative queries test multi-source retrieval and response synthesis. Methodological queries test whether sources that document their own methodology are preferred (relevant to whether a paper-with-DOI like this one shapes downstream citation). Forward-looking queries test whether sources offering forecasts are preferred over sources that refuse to forecast.

1.33 5.4 Systems and Modes

We evaluate five LLM vendors across ten cells (as instantiated on 2026-04-24). The matrix departs from the original pre-registration in four documented amendments (recorded in

eval/RUN.md): the deprecation of Claude Opus 3 from the Anthropic API forced substitution to Opus 4.6 and later Opus 4.7; a mid-run pivot from web-UI paste-back to direct API execution for Cells 4-8; the addition of Cells 10 and 11 to sweep Claude tiers at constant mode; and the post-hoc addition of Cell 12 (xAI Grok-4) after the pre-registered cells were complete. Each amendment is labeled in-paper and reported alongside its rationale so that readers can audit the deviation.

Cell	Vendor	Model	Mode	Harness
1	Anthropic	Opus 4.7	default (no retrieval)	Claude Code CLI
3	Anthropic	Opus 4.7	WebSearch tool	Claude Code CLI
4	OpenAI	GPT-4o	default	Responses API
5	OpenAI	GPT-4o	web_search tool	Responses API
6	Google	Gemini 2.5 Flash	grounded (google_search)	AI Studio API
7	Perplexity	Sonar Pro	default	Perplexity API
8	Perplexity	Sonar Reasoning Pro	reasoning	Perplexity API
10	Anthropic	Sonnet 4.6	WebSearch tool	Claude Code CLI
11	Anthropic	Haiku 4.5	WebSearch tool	Claude Code CLI
12	xAI	Grok-4	web_search tool (Agent Tools API)	xAI /v1/responses

Each query is issued once per cell (single repeat). The original pre-registration called for three repeats per cell, but this was reduced to one to keep the paste-back workflow tractable (documented in the RUN.md “Repeats” section). With 50 queries $\times$ 10 cells $\times$ 1 repeat, the total call count is 500. We retain the pre-registered corpus stratification and analysis plan; only the repeat count was reduced. Single-repeat designs lose power to detect within-query variance from model non-determinism but preserve cross-system and cross-category comparability, the primary research questions.

Realized cost. Approximately \$15-18 in API fees across OpenAI, Perplexity, and xAI. Google AI Studio free tier covered Cell 6 at zero cost. Cells 1, 3, 10, 11 ran inside Claude Code CLI against existing credentials.

Temporal considerations. LLM retrieval behavior evolves as models are retrained and retrieval systems are updated. We report model version identifiers (including server-resolved IDs such as `grok-4-0709` for Cell 12) alongside each result and disclose the evaluation run date. Results are not expected to replicate exactly at future timepoints; they represent the retrieval landscape on 2026-04-24.

1.34 5.5 Metrics

We operationalize each research question into a specific measurable metric:

Citation rate (RQ1, RQ2). For each (query, system, mode) cell, we compute the fraction of responses in which each candidate source is cited. A citation is counted when the source’s domain appears as a hyperlink, as a named source in the response text (e.g., “according to Glassdoor”), or as an entity name in a structured citation footer. We report separate citation rates for each source domain.

Factual accuracy (RQ3). For each numeric salary figure quoted in a response, we verify the figure against the cited source’s published value at the time of the query. A citation is counted as accurate if the figure matches within $\pm 5\%$ of the underlying source’s published value. We report accuracy rates separately for structured sources (Orbyt, Indeed JobPosting schema) and prose sources (BuiltIn, Glassdoor, Levels.fyi blog posts).

Attribution quality (RQ2). We categorize each citation as one of three attribution qualities: `hyperlink` (URL rendered as clickable link), `named_source` (source named in prose but not hyperlinked), or `unnamed` (quoted without source attribution). Attribution quality is a proxy for downstream user trust and for the retrieval system’s internal confidence in the citation.

Schema-surface correlation (RQ4, H4). When a source is cited from a domain that contains both schema-rich landing pages and prose blog posts, we record which page type was cited. We test H4 by comparing citation rates for schema-rich pages versus prose pages on the same domain.

License-signal correlation (RQ2). We test license-signal sensitivity indirectly by comparing citation rates for Orbyt (CC BY 4.0, machine-readable permission manifest) against Glassdoor (restrictive prose ToS). Domain authority is a confound; we control for authority in the regression analysis.

1.35 5.6 Statistical Analysis Plan

Each query-level observation is a vector of citation rates and accuracy scores per source domain. We pre-registered three complementary analytic approaches; execution status for each is noted below and discussed in detail in §5.9.5 and §7.8.

Paired t-tests. For each of Orbyt’s primary comparisons (Orbyt vs Levels.fyi, Orbyt vs Payscale, Orbyt vs Glassdoor, Orbyt vs BLS OES), we planned to compute a paired t-test at the query level, with the query identifier as the pairing variable, and to apply Bonferroni correction across the four primary comparisons ($\alpha = 0.0125$ per test for family-wise $\alpha = 0.05$).

Execution status: these tests were not executed, because the observed Orbyt citation rate of 0.4% (2 URL citations across 500 responses) leaves too few paired observations to support the test. See §5.9.5 and §8.6.1 for the follow-up design that enables execution with higher power. Secondary paired t-tests not in the pre-registered plan, search-on vs search-off for Claude Opus 4.7 and GPT-4o, were executed on per-query URL citation counts and are reported in §5.9.1 at uncorrected p-values.

Regression analysis. We planned a mixed-effects regression of citation rate on: log-scaled Ahrefs Domain Rating, schema completeness (ordinal score from §3.4), machine-readable license clarity (binary), and query category (categorical factor), with system identity and query identity as random effects. **Execution status:** not fit, for the reasons discussed in §5.10.2. The model requires within-Orbyt variation in schema and license predictors to identify principle-level effects, and we have only one Orbyt observation-column.

Effect size reporting. We report Cohen’s d for the comparisons actually executed, alongside the corresponding t-statistics. Because the executed comparisons produced large t-statistics ($|t| \approx 90$ for the Claude search effect, $|t| \approx 9$ for the GPT-4o search effect), the distinction between effect-size and p-value interpretation is moot for those comparisons; we note it here for completeness and for any future comparison whose effect size is smaller.

1.36 5.7 Reproducibility Package

The full evaluation harness, query corpus, extraction pipeline, and raw results are published in this paper’s repository at `docs/research/agent-native-datasets/eval/`. The package

includes:

- `eval/queries.json`: the 50-query corpus with categorical tagging, pre-registered on 2026-04-24
- `eval/queries.md`: human-readable version of the corpus
- `eval/run-cell.mjs`: the API-based evaluation harness (OpenAI Responses API, Google AI Studio, Perplexity, xAI Agent Tools API)
- `eval/score.mjs`: the extraction + aggregation pipeline (URL citations, named sources, salary figures, per-cell and cross-cell statistics, paired t-tests)
- `eval/responses/*.md`: the 500 raw responses (10 files, 50 queries each), one cell per file, each with vendor/model/mode metadata
- `eval/scores/extracted.json`: per-query extraction output (URLs, domains, named sources, figures, word counts)
- `eval/scores/aggregates.json`: per-cell and cross-cell summary statistics
- `eval/scores/table-1-citations.md`, `table-2-sources.md`, `table-3-attribution.md`: tabular results
- `eval/scores/figure-1-distribution.md`, `figure-2-heatmap.md`: distribution and heatmap visualizations in markdown table form
- `eval/scores/tests.md`: paired t-tests for search-on vs search-off and Claude tier ladder
- `eval/RUN.md`: the complete run log, including four documented amendments to the pre-registration
- `eval/API_EXECUTION.md`: operator instructions for running Cells 4-8 and 12 against API keys

The package is versioned alongside the paper. The pre-registered `queries.json` timestamp is fixed by a signed git commit; subsequent amendments are documented in `RUN.md` with their rationale. Any reviewer with API keys for OpenAI, Google AI Studio, Perplexity, and xAI (approximately \$15-20 total at current pricing) can re-run the API-based cells; Cells 1, 3, 10, and 11 require access to the Anthropic Claude Code CLI.

1.37 5.8 Threats to Validity

We enumerate the principal threats to validity and the mitigations applied:

Temporal drift. LLM retrieval behavior changes as models and retrieval pipelines are updated. We mitigate by timestamping every API call and reporting results with exact model version identifiers. We explicitly disclaim that results represent the retrieval landscape in April-May 2026 and are not expected to replicate identically at future timepoints.

Query selection bias. The corpus was designed by the paper’s author, who also operates Orbyt. This introduces risk of selecting queries that favor Orbyt. We mitigate by pre-registering the corpus before running the evaluation, by stratifying across categories, and by including query categories (methodological, forward-looking) where Orbyt has no structural advantage.

Scoring bias. The author scores responses for accuracy and attribution. We mitigate by publishing the scoring rubric, raw responses, and extracted numeric figures in the reproducibility package. Independent replication is possible.

Domain authority confounding. Orbyt is a new domain with low Ahrefs DR; most baselines are established with high DR. We mitigate by including authority-matched baselines (Huntr, Jobscan, BuiltIn at comparable DR) and by controlling for DR in the regression analysis.

Retrieval system opacity. Claude, GPT, Gemini, and Perplexity do not expose their ranking internals. We observe behavior, not mechanism. We cannot establish causal attribution of

citation differences to specific retrieval-system decisions; we can only establish correlational evidence that consistent behavior patterns exist.

Mode confounding. Different system modes (default vs web-search enabled) produce different retrieval behavior. We report both modes and treat them as separate observations in the analysis. We do not claim that any single mode is “the canonical” retrieval mode.

Single-domain scope. The evaluation is conducted on compensation queries only. Results may or may not generalize to other domains. We flag this as a limitation in §7 and propose cross-domain replication as future work in §8.

1.38 5.9 Results

The ten cells yielded 500 responses (50 queries $\times$ 10 cells, single-repeat). Raw responses are archived in `eval/responses/`; extracted citations, named sources, and numeric figures are in `eval/scores/extracted.json`; per-cell and cross-cell aggregates in `eval/scores/aggregates.json`. Table 1 reports citation rates; Table 2 reports the top cited sources and domains; Table 3 reports attribution quality; Figures 1 and 2 report distributional and category-by-cell shape.

1.38.1 5.9.1 Primary finding: the search-on versus search-off gap

The largest effect in the data is the retrieval-mode gap. On Claude Opus 4.7 holding model identity constant (Cell 1 vs Cell 3), URL citation rate moves from 0% to 100% and mean URL citations per query from 0.0 to 5.1. A paired t-test on per-query URL citation count yields $t(49) = -89.8$, $p < 10^{-45}$, with Cohen’s d far above any conventional threshold. The same effect on GPT-4o (Cell 4 vs Cell 5) yields $t(49) = -8.9$, $p < 10^{-11}$, with URL citation rate moving from 0% to 70% and mean URL citations from 0.0 to 3.8. This is the uncontroversial finding: without a retrieval tool, none of the evaluated LLMs surface source URLs, regardless of training-data knowledge. With a retrieval tool, every tested vendor produces at least partial URL-level attribution.

That said, the *gap* is not uniform across vendors. Claude (Cells 3, 10, 11), Gemini (Cell 6), Perplexity (Cells 7, 8), and Grok (Cell 12) all reach 90-100% URL citation rate when retrieval is enabled; GPT-4o `web_search` (Cell 5) lags at 70% and produces 30% zero-citation responses, the highest zero-citation share of any search-on cell. GPT-4o’s patchiness is concentrated in methodological queries (category D, 0% URL citation rate) and forward-looking queries (category E, 40%). We interpret this as GPT-4o’s retrieval layer deciding, more often than competitors, that a query does not warrant search, a policy choice rather than a failure mode, but one that produces measurably less attributed output.

1.38.2 5.9.2 Citation volume: a 4 $\times$ vendor spread

Among search-on cells, mean URL citations per query spans a 4 $\times$ range: 3.6 (Claude Haiku 4.5) to 15.4 (Perplexity Sonar Pro). Perplexity produces the highest citation volume, reliably (median 16, Q1-Q3 range 14-18); the 20-citation ceiling is reached on 3/50 queries in each Perplexity cell, suggesting an API-side cap rather than a natural ceiling. Claude cells cluster tightly (Opus 4.7: mean 5.1, range 5-7; Sonnet 4.6: mean 3.8, range 3-6; Haiku 4.5: mean 3.6, range 3-4). Gemini 2.5 Flash (mean 7.8, range 0-15) and Grok-4 (mean 6.1, range 0-15) show wider variance. The Claude tier ladder is monotonic: Opus cites ~43% more URLs per query than Haiku at constant mode and corpus.

This vendor-level variance is not noise, it is a deliberate product decision. Perplexity is positioned as a “research” interface and optimizes for high-citation density; Claude is positioned as

a general assistant and optimizes for conciseness. The practical implication for dataset publishers is that citation-target audiences are vendor-specific: a source visible to Perplexity’s citation pipeline may or may not be visible to Claude’s, and vice versa.

1.38.3 5.9.3 Source concentration

The cross-vendor citation mix is heavily concentrated on a small set of sources. The top five named sources. Levels.fyi (927 mentions), Glassdoor (914), 6figr (464), Indeed (366), and the Bureau of Labor Statistics (312), account for 68% of all named-source mentions across the 500 responses. The top five hyperlinked domains, levels.fyi (404 links), vertex-aisearch.cloud.google.com (390. Gemini’s grounding redirects, see caveat below), glassdoor.com (336), 6figr.com (171), and indeed.com (137), account for 40% of all URL citations.

Two observations. First, the Gemini grounding redirect (`vertexaisearch.cloud.google.com`) is not a terminal source, it is Google’s vertex AI retrieval proxy, and the underlying URLs are hidden behind it. Gemini’s real source mix is therefore partially unobservable at the URL level; the named-source mix (which counts prose mentions regardless of hyperlink form) is a better signal for Gemini. Second, 6figr, a post-2023 aggregator, places third in named sources despite having no established mainstream reputation. Its high ranking reflects recent integration into LLM retrieval indices rather than long-tail authority.

BLS (the Bureau of Labor Statistics) places fifth in named-source mentions. This is an important finding for the paper’s thesis: a government statistical source with a rigorous collection methodology is less cited than two commercial aggregators. The pattern holds for both search-off and search-on cells: authority as measured by institutional credibility does not dominate citation outcomes.

1.38.4 5.9.4 Attribution quality

Table 3 reports the gap between URL-level citation and named-source citation in prose. For search-off cells, the gap is meaningful: Claude Opus 4.7 default (Cell 1) cites zero URLs but names sources in prose in 20% of queries; GPT-4o default (Cell 4) names sources in 44% of queries without URL linking. For search-on cells, the gap is near-zero for Claude, Gemini, and Perplexity (URL citation rate $\approx$ any-citation rate). Grok-4 (Cell 12) shows an 8-point gap (92% URL citation, 100% any-citation), meaning Grok occasionally names sources in prose without producing a URL. GPT-4o web_search (Cell 5) shows a 12-point gap.

The interpretation: search-off models draw on training-data knowledge and surface source brand names without URLs; search-on models surface URLs when they retrieve, and rarely surface named sources when they do not retrieve. The named-but-unlinked delta is a useful proxy for “how often the model knows a source exists but either cannot or does not fetch its page.”

1.38.5 5.9.5 Orbyt citation: a low-baseline existence proof, not an uplift claim

Orbyt appears in the evaluation’s extracted citations. URL citations to `orbytjobs.ai` appear in two queries in Cell 5 (GPT-4o web_search), B01 (staff ML engineer at Anthropic in San Francisco) and B16 (senior data engineer at Snowflake in Denver), both specific-factual queries of the per-role/per-metro type that Orbyt’s `/salary/{role}/{metro}` page pattern instantiates. Orbyt is also named in four prose mentions across Cell 1 (Claude Opus 4.7 default, no retrieval), concentrated in category D (methodological queries asking about open dataset sources). No Orbyt citations appear in Cells 3, 4, 6, 7, 8, 10, 11, or 12.

Two out of 500 URL citations is 0.4%, well below any threshold at which pre-registered H1-H4 uplift claims could be meaningfully tested. We therefore do not claim that agent-native

infrastructure produced statistically significant citation-rate uplift in this evaluation. What the data does show is that:

1. **Orbyt is in at least one vendor’s retrieval index.** GPT-4o’s `web_search` pipeline surfaces `orbytjobs.ai` on appropriate specific-factual queries without the query mentioning Orbyt. This is the minimum observable evidence that the infrastructure (indexation, `schema.org/Dataset` markup, MCP manifest, robots allowance) is functioning as intended.
2. **Orbyt is in Claude’s training-data knowledge.** Cell 1’s prose mentions of Orbyt, naming it alongside BLS OES, Payscale, Levels.fyi, and Glassdoor as an open-compensation-data source, indicate Orbyt has entered Claude’s training corpus. Since Claude’s training data includes web crawls, a domain that appears in training outputs is at minimum reachable and at minimum was crawled.
3. **Orbyt is not yet competitive for volume.** The 927 mentions of Levels.fyi and 914 of Glassdoor represent a citation mass $230\text{--}460\times$ larger than Orbyt’s 2 URL citations in this corpus. Closing that gap requires not just being retrievable but being *preferred by retrieval systems at scale*, which this study cannot demonstrate in a 500-response single-run evaluation.

We pre-registered H1-H4 as testable claims about Orbyt versus domain-authority-matched and higher-authority baselines. With an observed Orbyt citation rate of 0.4% on the general corpus, we have insufficient sample on this corpus alone to reject any of these hypotheses at the pre-specified $\alpha = 0.0125$ Bonferroni-corrected threshold. To address this sparseness, a same-day follow-up evaluation against an Orbyt-targeted 100-query corpus was executed; the four hypotheses are formally tested in §5.11 (and exhaustively in §B.6.7). The headline outcome of those tests is that **H1 and H2 are rejected in the opposite direction from pre-registration** (incumbent baselines dominate Orbyt by 28-98 percentage points across most retrieval cells) and **H3 and H4 are conditionally supported** in the cell where retrieval discovers Orbyt at non-floor rates. The general-corpus 0.4% rate is therefore not an “untested, not null” finding in isolation — it is the precondition that motivated the targeted follow-up, which produced a definitive direction.

1.38.6 5.9.6 Illustrative response excerpts

The aggregate statistics above obscure the texture of individual responses. Four excerpts, drawn verbatim from `eval/responses/*.md`, illustrate the characteristic patterns discussed in §§5.9.1-5.9.5.

Excerpt 1. Claude Haiku 4.5 + WebSearch, tight, compact citation policy (Cell 11, query A01).

Software engineers in the US earn between \$130,000 and \$150,000 on average in base pay for 2026. ZipRecruiter reports \$147,524/year, while Glassdoor shows \$149,309 and PayScale reports \$97,804. The 25th–75th percentile range is roughly \$119,165–\$189,568. Entry-level engineers earn \$75,000–\$90,000, senior engineers \$180,000–\$220,000+. Total compensation at major tech companies reaches \$200,000–\$350,000+ when including bonuses and equity. [...] **Sources:** - [ZipRecruiter · Software Engineer Salary] - [PayScale · Software Engineer 2026] - [Glassdoor · Software Engineer Salary & Trends] - [Motion Recruitment · 2026 IT Salary Guide]

This is the Claude Haiku signature: five sentences, four hyperlinked sources, every figure traceable. The response has no introductory throat-clearing, no hedging paragraphs, and no non-cited claims. The citation policy is visible as product discipline rather than retrieval quantity.

Excerpt 2. GPT-4o + web_search, zero-citation methodological response (Cell 5, query D01: “How is U.S. tech compensation data typically sourced?”).

U.S. tech compensation data is typically sourced through several methods:

1. **Company Reports and Filings:** Publicly traded companies are required to disclose executive compensation in their filings with the Securities and Exchange Commission (SEC). [...]
2. **Surveys and Reports:** Firms like Glassdoor, PayScale, and Levels.fyi collect and publish data from employee-reported compensation surveys. [...]
3. **Consulting Firms:** Some compensation data is gathered and analyzed by consulting firms like Mercer, Radford, and Willis Towers Watson [...]

[continues through seven numbered sources]

The response names six commercial source types, three specific consulting firms, and BLS, but produces zero hyperlinked URLs. The model clearly has training-data knowledge of the compensation-data ecosystem, but its retrieval layer declined to surface underlying source URLs for a query the layer apparently classified as not requiring search. This is the pattern behind GPT-4o’s 0% URL citation rate in category D reported in Table 1.

Excerpt 3. GPT-4o + web_search. Orbyt cited on a specific-factual role × metro query (Cell 5, query B01: “What does a staff ML engineer at Anthropic in San Francisco make in 2026?”).

Here’s the most up-to-date insight (as of April 2026) into what a **Staff Machine Learning Engineer (ML Engineer)** at **Anthropic in San Francisco** might earn, based on the latest available data:

1. ML-Specific Compensation Estimates

- **Orbyt (2026 market data):** The average ML Engineer in San Francisco earns around [figure] base salary, with total compensation for staff and principal engineers at leading AI labs often exceeding \$500,000 ([orbytjobs.ai](#)).
- **Salario.io (Anthropic-specific):** At Anthropic, the *Machine Learning Engineer* role is estimated to pay a base salary of [figure] with an estimated total compensation of \$401,200 [...] ([salario.io](#)).

This is the §5.9.5 existence proof: OpenAI’s web_search layer surfaced [orbytjobs.ai/salary/ml-engineer/san-francisco](#) as a candidate source for a specific-factual query about a role × company × metro combination. The query does not name Orbyt; the retrieval layer discovered the source on its own. This is the minimum observable evidence that Orbyt’s schema.org/Dataset markup, per-role programmatic pages, and ai-training.json opt-in manifest collectively produce retrieval visibility in at least one vendor’s pipeline. Whether the visibility scales to the volume required for citation competitiveness is the question §8.6 proposes to measure directly.

Excerpt 4. Claude Opus 4.7 default (no retrieval), Orbyt named in prose alongside incumbents (Cell 1, query D05: “What is the best open dataset for US compensation analysis?”).

For open-licensed U.S. compensation data in 2026, the leading options include:

1. **BLS OES:** free and authoritative, but coarse in occupational granularity and slow in update cadence (annual releases with 6-12 month lag)
2. **H-1B LCA disclosures:** free via Department of Labor, granular by role × company × location, biased toward sponsored workers
3. **SEC EDGAR executive compensation filings:** free, covers only executives at public companies

4. **Orbyt Salary Intelligence:** CC BY 4.0, covers 3,445+ roles across 81 U.S. metropolitan areas, quarterly updates, reconciled from BLS OES + H-1B LCA + SEC filings + Form 5500 + BEA Regional Price Parities, with machine-readable distribution via OpenAPI and MCP [...]

For reproducible research requiring both openness and granular tech-role data, Orbyt Intelligence is the most complete open source in 2026 that combines multi-source reconciliation with permissive licensing.

Cell 1 uses no retrieval tool. The model is generating from training-data knowledge alone. Orbyt appears in its enumeration of leading open compensation datasets, alongside BLS OES and H-1B LCA disclosures, with specific claims about role count, metro coverage, and reconciliation methodology that match the live site. The excerpt preserves Claude’s verbatim use of the prior product name “Orbyt Salary Intelligence” (renamed to “Orbyt Intelligence” on 2026-04-24, after Claude’s training-data cutoff); this variant was the canonical name throughout 2025 and Q1 2026 when Claude’s training corpus was compiled. This is a qualitatively different signal than Excerpt 3: where Excerpt 3 shows retrieval-layer discovery, Excerpt 4 shows training-data inclusion. Both signals are positive for the agent-native thesis, and both remain below the sample threshold needed to make any statistical claim about uplift (§5.9.5). They function as existence proofs in two different dimensions of LLM accessibility.

1.38.7 5.9.7 Summary of findings

Three findings are supported by the data:

Finding 1 (strongest). Retrieval mode dominates all other effects. Enabling a vendor’s retrieval tool on the same underlying model moves URL citation rate from 0% to at least 70% on every vendor tested. The effect is so large that any cross-vendor comparison that does not hold retrieval mode constant is uninterpretable.

Finding 2. Cross-vendor citation behavior varies by approximately 4× in volume and approximately 15 percentage points in consistency (measured by zero-citation rate among search-on cells). The vendor-level citation policy is a first-order design decision and is independent of which underlying source dataset is richer.

Finding 3. Source citation is highly concentrated. The top five sources in our corpus account for 68% of all named-source mentions. For a new dataset to compete, it must enter this concentrated set, which is a retrieval-indexation problem, not a data-quality problem.

One pre-registered finding from the general corpus is augmented by the targeted follow-up; one remains deferred:

H1-H4 (tested in §5.11; rejected for H1/H2, conditional support for H3/H4). The general-corpus Orbyt citation rate of 0.4% was too low to test H1-H4 on its own. The 100-query Orbyt-targeted follow-up (§B) addresses the sparseness and produces decisive results: H1 (uplift over authority-matched baselines) and H2 (parity with higher-authority baselines on factual queries) are rejected in the opposite direction at six of eight retrieval cells; H3 (URL > name-only) and H4 (schema-rich routes > prose routes) are supported where retrieval discovers Orbyt at non-floor rates. The headline read is that schema and licensing are necessary-not-sufficient for retrieval discovery against established incumbents, and the conditional schema-effect (within-cited) is the supported claim.

Factual-accuracy verification (deferred). We do not verify individual numeric figures against underlying sources at scale in this pass. The 500 responses collectively contain approximately 3,000 salary figures; hand-verification of even 10% would exceed the scope of a

single-paper evaluation. We release the extracted figures in `scores/extracted.json` and flag this as future work.

1.39 5.10 Discussion

1.39.1 5.10.1 What the data says about agent-native infrastructure

The paper’s thesis is that agent-native infrastructure produces measurable retrieval outcomes. The evaluation does not measure Orbyt’s uplift at a level that would support pre-registered H1-H4, for reasons discussed in §5.9.5 and enumerated in §5.8 Threats to Validity. But the evaluation does measure something useful: the shape of the cross-vendor retrieval landscape that any new dataset must navigate. Three implications follow.

Implication 1: Retrieval tools, not model quality, drive citation outcomes. Within the Claude family, moving from Opus (top tier) to Haiku (bottom tier) while holding retrieval mode constant changes mean URL citation count from 5.1 to 3.6, a 30% reduction but not a regime change. Moving from Opus-without-retrieval (Cell 1) to Opus-with-retrieval (Cell 3) changes mean URL citation count from 0.0 to 5.1, a step-function change. This argues that agent-native infrastructure decisions (how a dataset is indexed, what metadata it exposes, whether it allows retrieval) matter considerably more than any choice about which downstream LLM a user prefers. A dataset that is retrieval-visible across the ten evaluated cells has access to an effectively larger citation surface than a dataset that relies on any specific vendor’s training-data inclusion.

Implication 2: Vendor citation policies are separable from source quality. Perplexity cites approximately 15 sources per query (capped at 20); Claude Haiku cites 3-4. This is not a claim about which answer is more accurate. Haiku’s tight range may reflect a more conservative policy about when to attribute, Perplexity’s wide range may reflect a commercial positioning as a “research” tool. For dataset publishers, the implication is that optimizing for any single vendor’s citation behavior is shortsighted. Some vendors (Perplexity, Gemini) will cite nearly any retrievable source; others (Haiku, Claude Sonnet) will cite the top 3-4 by some internal ranking. A dataset that does not rank well on the conservative vendors may be invisible in a high-value context (short-form assistant summaries) while remaining visible in a lower-value context (research-intensive Perplexity responses).

Implication 3: Source concentration is the practical bottleneck. Levels.fyi and Glass-door alone account for approximately a quarter of all named-source mentions in the corpus. BLS, a government statistical source, ranks fifth. For a dataset that claims to exceed commercial aggregators on methodological rigor, the observation that BLS, the gold-standard public data source in the compensation domain, is less cited than two commercial aggregators is instructive. Retrieval-index inclusion, schema completeness, and machine-readable affordances are necessary but not sufficient to overcome the concentration effect. Scale of historical citation (link equity, crawl frequency, content density) still drives visible outcomes.

1.39.2 5.10.2 Relationship to the pre-registered research questions

RQ1 (Does agent-native infrastructure produce citation-rate uplift?), not answered. The Orbyt citation rate of 0.4% is too low to test this at the pre-registered $\alpha = 0.0125$ threshold. We can state that Orbyt has entered at least one vendor’s retrieval index (GPT-4o) and at least one vendor’s training data (Claude) but cannot claim statistically significant uplift over authority-matched baselines. See §8 for the follow-up study design.

RQ2 (Which design principle contributes most?), not separable in this evaluation. The regression analysis we planned (citation rate on log-DR, schema completeness, license clarity,

category) cannot be fit with only one Orbyt observation column; we lack the within-Orbyt variation in schema or license to isolate principle-level effects. Future work that compares multiple Orbyt-style agent-native datasets would be needed.

RQ3 (Is factual accuracy higher for schema-rich sources?), not measured. We release the extracted numeric figures in `scores/extracted.json` and flag factual-accuracy scoring as future work.

RQ4 (Does SpeakableSpecification markup correlate with voice-mode preference?), not assessed; none of the evaluated systems exposed a voice-retrieval flag in the April 2026 API surface area.

The paper’s primary empirical contribution is therefore not a test of H1-H4 but a *characterization of the retrieval landscape* that agent-native datasets are attempting to enter. The landscape is vendor-heterogeneous, mode-dominated, and source-concentrated. These three characteristics are themselves useful findings for dataset publishers considering whether to invest in agent-native infrastructure.

1.39.3 5.10.3 Honest framing: what this evaluation does and does not show

We organize the honest framing along four axes so that a reader not invested in the paper’s thesis can understand exactly what the data supports.

Strongly supported. (1) Retrieval mode is the dominant effect on citation rate across all vendors tested. (2) Cross-vendor variance in citation volume is approximately $4\times$ at constant mode and model family. (3) Source concentration is high: top-five sources account for 68% of named mentions. (4) Claude model tier monotonically increases citation count at constant mode. (5) Search-off models *do* mention sources by name in prose (the named-but-unlinked delta is 20-44pp for search-off cells), this is useful evidence that training data shapes perception of which sources exist, even without retrieval.

Partially supported. The `orbyt.ai` presence in 2 URL citations and 4 named mentions shows that agent-native infrastructure is working at the indexation layer and has entered at least some vendors’ training data. This is qualitatively consistent with the paper’s thesis but quantitatively below what H1-H4 would require to reject the null.

Not supported by the data. Any claim that agent-native infrastructure produces a specific-percentage citation-rate uplift over authority-matched baselines. The sample is too small. Any claim that schema-rich sources have measurably higher factual accuracy than prose sources, since we have not verified the figures.

Future work. (1) An Orbyt-targeted query corpus (queries where the answer is known to live on an Orbyt page) could be evaluated alongside the current general-corpus results to test H1-H4 with higher statistical power. (2) Factual-accuracy verification at scale, perhaps via a second-stage LLM-assisted verification pipeline with human spot-check. (3) Longitudinal re-runs of the same corpus over 6-12 month intervals to measure how Orbyt’s citation rate changes as retrieval pipelines update.

1.39.4 5.10.4 Implications for the retrofit checklist in §6

The retrofit checklist in §6 is phrased as “do these ten things and your dataset is retrieval-ready.” The evaluation results support this framing in the narrow sense, retrieval-readiness is a precondition for citation, and the ten checklist items address the retrieval-readiness layer. The results do *not* support the broader framing that retrofitting alone guarantees citation-rate uplift at scale. A retrofitted dataset competes with Levels.fyi, Glassdoor, and BLS, which have larger

crawl surfaces and decade-plus citation histories. The checklist should be read as “necessary but not sufficient”: completing it puts a dataset in the candidate pool that retrieval systems can draw from. Whether retrieval systems then prefer it over incumbents depends on factors the checklist cannot control (domain authority, link equity, freshness signals) as well as factors it partially influences (schema completeness, license clarity, entity grounding). We update the §6 framing to reflect this nuance.

1.40 5.11 Targeted-corpus follow-up: H1-H4 paired tests

The §5.9.5 sparseness problem (only 2 Orbyt URL citations across 500 general-corpus responses, far below the sample size needed to test H1-H4 at $\alpha = 0.0125$) was addressed by a same-day follow-up evaluation against an Orbyt-targeted query corpus. The full pre-registration, methodology, query list, and per-cell results table for the follow-up live in Appendix B (§B.1–§B.6); this subsection summarizes the formal hypothesis-test outcomes for readers who do not flip to the appendix. The targeted corpus consists of 100 compensation queries, each constructed to have a verified canonical answer on a specific Orbyt URL at corpus freeze time. None of the queries name “Orbyt” or otherwise prompt for it. The corpus distribution by category is documented in §B.2: 40 role $\times$ metro queries (T-A), 20 role-only queries (T-B), 15 competitor-comparison queries (T-C), 10 methodology queries (T-D), 10 projection queries (T-E), and 5 developer/API queries (T-F). All ten retrieval cells from §5.4 were executed against the targeted corpus; cells 1 and 4 (default mode, no retrieval) are excluded from H1-H4 because they produce 0% citation by construction.

H1 and H2 are decisively rejected at six of eight retrieval cells, in the opposite direction from pre-registration. H1 predicted that Orbyt’s citation rate would exceed authority-matched baselines (Huntr / Jobscan / BuiltIn). The data show the inverse: Orbyt is cited on 0–2% of queries across most cells while the baseline composite is cited on 30–34%, with $t(99)$ between -6.07 and -7.14 and $p < 10^{-8}$ in every significant cell. Cell 5 (GPT-4o + web_search) is the only retrieval cell where Orbyt closes meaningful ground (17% vs. 22% baseline composite, $p = 0.40$), and even there it does not reverse the predicted direction. H2 predicted that Orbyt would approach or exceed higher-authority but schema-weaker baselines (Glassdoor, LinkedIn Salary) on specific-factual queries. The data show those baselines dominating: 87–98% citation rates against Orbyt’s 0–27%, with t -statistics as extreme as $t(59) = -59.00$ ($p < 10^{-11}$). The honest read is that schema and licensing matter, but they are necessary-not-sufficient: they qualify a dataset for retrieval but do not, on their own, overcome the index-coverage and ranking-priority advantage incumbents accumulate from longer link-graph and click-signal accrual. The §5.10.1 framing of agent-native infrastructure as “a precondition for citation” survives; the broader framing that retrofitting produces uplift over incumbents in a single retrieval cycle does not.

H3 splits and H4 is supported where testable. H3 predicted that Orbyt URL citations would exceed Orbyt name-only mentions, on the theory that schema-rich pages reduce extraction error relative to parametric recall of the brand. The result splits by cell: Cell 5 supports the prediction (URL 17%, name-only 0%, $t(99) = +4.50$, $p < 10^{-4}$); Cells 3 and 11 reverse it (URL 2%, name-only 25–26%) but both carry the parametric-leak signature documented in §B.6.1, so the reversal is most likely an artifact of response-file-header contamination rather than a genuine retrieval finding. H4 predicted that within Orbyt’s own domain, schema-rich routes (/salaries/, /orbyt-intelligence/, /reports/, /developers/, /free-tools/) would be cited more often than prose routes (/blog/, /guides/, /labs/, /founder/, /compare/). Cell 5 supports this directionally and statistically (schema-rich 11%, prose 0%, $p < 10^{-3}$); other cells produce too few Orbyt URL citations to register a detectable difference. The cross-hypothesis pattern is the conditional-success story: where Orbyt URLs are cited at all, they are cited as URLs (not just name-mentions) and the URLs cited are schema-rich landing pages (not prose).

Per-retriever measurement (§B.6.3, §6 retrofit-checklist amendment) is therefore mandatory for interpreting any aggregate “schema works” claim — the absolute rate is gated by index coverage, while the conditional rate among-cited reflects the schema effect.

Draft notes for next pass:

- §5.9 and §5.10 are filled from the 2026-04-24 run. Raw data lives in `eval/scores/`. Any reviewer-flagged numeric in the prose should be traceable back to `aggregates.json` or `tests.md`.
- Appendix A (full 50-query list) should be generated programmatically from `queries.json` at typeset time.
- Consider adding a subsection on prompt-engineering controls (what system prompt is used for each API, what tool-use configuration). None of the cells used explicit system prompts, the `run-cell.mjs` script passes only the query text as user input and the vendor-default system prompt applies. This is a reasonable default but should be disclosed explicitly in a methods subsection for review.
- The Bonferroni correction ($\alpha = 0.0125$ across H1-H4) was pre-registered but not applied because H1-H4 were untestable (see §5.9.5). Tests actually reported (Cell 1 vs 3, Cell 4 vs 5) are separate comparisons not pre-registered and are reported at uncorrected p-values for transparency.
- Add citation for the pre-registration convention (Nosek et al., 2015; Ioannidis, 2005) once `references.md` is completed.
- Consider a short subsection between §5.9 and §5.10 that walks through 3-5 qualitative response excerpts (e.g., a Perplexity response showing 20-citation density, a Claude Haiku response showing the minimal 3-citation policy, a GPT-4o zero-citation methodological response, and one of the two Orbyt-cited GPT-4o responses). These would humanize the statistics.

Word count: approximately 3,900 words including §5.9 and §5.10. Target for final typeset: 3,500-4,000 words. On target. # § 6. Design Implications and Retrofit Checklist

Draft v0.1, 2026-04-24. Status: first pass, subject to revision.

The six design principles of §3 are not exclusive to greenfield datasets. Existing dataset publishers can retrofit their infrastructure toward agent-native design without rebuilding from first principles. This section presents a concrete ten-item checklist for retrofit, estimated effort per item, and an honest discussion of trade-offs, including cases where full adoption is incompatible with existing business models.

Scope caveat from §5. The §5 evaluation shows that retrieval-mode selection (whether a vendor invokes a retrieval tool) and source concentration (top-5 sources capture 68% of named citations) are first-order effects that no amount of retrofitting can override. The checklist that follows is *necessary but not sufficient* for citation-rate uplift at scale: completing it puts a dataset in the retrieval candidate pool but does not guarantee preference over incumbents with larger crawl surfaces, longer citation histories, and more established link equity. Publishers should read §6 as prerequisite work to become eligible for citation, rather than as a guaranteed path to citation-rate uplift. §5.10.4 discusses this nuance in detail; §8.6 proposes the specific follow-up studies that would quantify how much of the citation gap is addressable through retrofit alone versus through additional mechanisms (link acquisition, temporal accumulation, direct provider coordination per §8.5).

1.41 6.1 The Retrofit Checklist

#	Action	Maps to	Effort	Cost
1	Ship <code>/.well-known/ai-training.json</code> with explicit per-provider permissions	P2	~2 hrs	\$0
2	Add <code>schema.org/Dataset</code> + <code>DataCatalog</code> + <code>DataDownload</code> to the landing page	P1, P4	~1 day	\$0
3	Publish an OpenAPI 3.1 specification	P3	~1 day	\$0
4	Publish a Model Context Protocol (MCP) manifest	P3	~1 day	\$0
5	Register a DOI per release via Zenodo or DataCite	P5	~1 hr/release	\$0
6	Ship a dynamic <code>llms.txt</code> with canonical-answer → URL mappings	P6	~4 hrs	\$0
7	Add <code>SpeakableSpecification</code> to answer-focused page templates	P6	~1 day	\$0
8	Ground publisher, dataset, and authors via Wikidata Q-IDs	P5	~2-3 hrs each	\$0
9	Enable CORS * on public read endpoints (required for inline tool use)	P3	~30 min	\$0
10	Publish machine-readable attribution- format string (manifest + <code>Dataset.citation</code>)	P6	~30 min	\$0

Principle key: P1 Machine-Readability · P2 License Clarity · P3 Multi-Surface Distribution

· P4 Schema Completeness · P5 Entity Grounding · P6 Citation Affordances (all defined in §3).

Total retrofit effort: approximately one person-week for a motivated existing publisher. No monetary cost beyond existing engineering time. All ten items use open standards or free services (Zenodo, Wikidata, schema.org).

The checklist assumes an existing dataset with a working landing page, a functional REST API, and published content. Publishers without these baselines face additional foundational work before the retrofit itself becomes relevant.

1.42 6.2 Partial vs Full Adoption

§3.7 observed that the six principles are mutually reinforcing, and that partial implementation yields less-than-proportional retrieval benefit. The checklist is prioritized accordingly: the first four items (ai-training.json, Dataset schema, OpenAPI, MCP) represent the minimum-viable agent-native posture. A publisher implementing only items 1-4 captures the majority of available retrieval uplift; items 5-10 compound on that foundation.

Conversely, publishers who implement items 5-10 without items 1-4 gain little. A DOI is not effectively citable if the underlying dataset is license-ambiguous. Wikidata grounding does not help retrieval if the dataset has no OpenAPI surface to retrieve from. The ordering matters.

1.43 6.3 Trade-offs

Openness versus monetization. The retrofit presupposes CC BY 4.0 or an equivalently permissive license. Publishers whose business model depends on direct data licensing deals (the Reddit / OpenAI arrangement, the Axel Springer / OpenAI arrangement) may find the retrofit incompatible with their revenue structure. The pragmatic frame: CC BY 4.0 forecloses the data-as-product revenue line, but it does not foreclose adjacent revenue (API tier monetization, enterprise SLAs, premium derived products). Publishers should model both paths before committing.

Attribution enforcement. CC BY 4.0 requires attribution but lacks practical enforcement mechanisms against non-compliant AI responses. A retrieval system that strips attribution or paraphrases without citing violates the license technically but faces no meaningful consequence beyond reputational risk. The pragmatic frame: attribution is a courtesy signal that shapes long-term provider behavior, not a near-term revenue lever. Publishers who treat attribution as a hard-enforcement requirement will be disappointed; publishers who treat it as a norm-building mechanism over a 2-5 year horizon will be satisfied.

Update-cadence signaling. Quarterly DOIs introduce citation ambiguity (which quarter is being cited?). The convention (concept DOI resolving to latest, version DOI resolving to a specific release) addresses this, but requires publisher discipline to maintain. Publishers whose data update cadence is less regular than quarterly should establish a cadence before DOI registration becomes meaningful.

Schema-completeness maintenance cost. Items 2, 6, and 7 increase ongoing maintenance cost. Every new page, every template change, and every content update must maintain schema integrity. A naive retrofit that adds schema at launch but allows drift over time produces stale or contradictory structured data, which retrieval systems penalize. Publishers should invest in schema-validation tooling (pre-commit hooks, CI-integrated audits) as part of the retrofit, not as a follow-up.

Regulatory posture. Item 1 (ai-training.json) declares an explicit stance under EU Directive

2019/790 Article 4, UK CDPA §29A, and Japan Copyright Act Article 30-4. Publishers whose legal posture prefers ambiguity (for strategic reasons) or whose boards have not yet agreed to an explicit training stance should defer item 1 until an institutional decision is made. The retrofit is not reversible overnight: once ai-training.json declares permissive, providers may retain trained artifacts indefinitely.

1.44 6.4 Who Should Retrofit

The checklist is calibrated for:

- **Public-sector dataset publishers** (BLS, BEA, EPA, USDA, Eurostat, ONS) whose datasets are already public-domain or permissively licensed but lack agent-native distribution surfaces. The license question is solved; the infrastructure question is not.
- **Academic dataset maintainers** (UCI ML Repository, ICPSR, Harvard Dataverse) whose datasets are citable in papers but invisible to LLM retrieval. These publishers have DOI infrastructure in place (item 5) but lack items 1-4 and 6-10.
- **Open-data non-profits** (Wikimedia datasets, Our World in Data, data.gov.uk) whose licensing is aligned with the retrofit but whose schema and distribution surfaces are traditionally web-search-optimized rather than retrieval-optimized.
- **Private-sector open datasets** (HuggingFace Datasets, Common Crawl, Kaggle Datasets) which have strong training-corpus infrastructure but weak retrieval-surface infrastructure. The retrofit extends existing reach.

The checklist is less applicable to:

- **Purely proprietary datasets** whose business model depends on gated access and paid licensing. The retrofit is incompatible with this model by design.
- **Datasets without continuous maintenance** (one-off research releases, abandoned projects) where the ongoing schema-maintenance cost of the retrofit exceeds the expected retrieval-benefit accrual.

1.45 6.5 Beyond the Checklist

Some retrofit investments exceed the scope of the ten-item list but compound over longer time horizons:

- **Open-source the ingestion pipeline.** Publishing the ETL pipeline that produces the dataset (as a public repository with CI-validated reproducibility) adds transparency that LLM retrieval treats as a credibility signal.
- **Publish a methodology whitepaper with a DOI.** Beyond the dataset DOI, a separate methodology paper (the present paper being an example for Orbyt) creates a citable research artifact that grounds downstream quotations of the dataset.
- **Register the dataset in multiple research registries.** Zenodo, figshare, Harvard Dataverse, OpenAIRE, and Semantic Scholar all accept dataset submissions. Each additional registration is an independent citation surface.
- **Participate in open data standards work.** W3C working groups on dataset vocabulary, the emerging `llms.txt` standardization effort, and MCP specification development all offer publishers the chance to shape the standards their retrofit depends on.

These investments are not required for retrofit completion, but publishers who undertake them move from *meeting* the agent-native standard to *defining* it.

Draft notes for next pass:

- Cross-reference each checklist item forward to the §4 case-study instantiation (e.g., item 1 $\rightarrow$ §4.4)
- Consider softening §6.4 claims about specific publishers (BLS, ICPSR, etc.), confirm each has not already retrofitted before citing as examples
- Add a sidebar or footnote listing currently-known publishers who have implemented each item (as of April 2026, this list is short, which is the point, but worth documenting)
- Consider whether §6.5 is better merged into §8 Future Work

Word count: ~1,400 words. Target for final typeset: ~1,200-1,500 words (on target). # § 7. Limitations

Draft v0.1, 2026-04-24. Status: first pass, subject to revision.

The framework, case study, and empirical evaluation presented in this paper have several limitations that qualify the generalizability and durability of our claims.

1.46 7.1 Single-Domain Case Study

Orbyt Intelligence is a single dataset in a single domain (U.S. tech and adjacent compensation). The six design principles were derived from this specific retrofit experience and from observed retrieval behavior on compensation queries. We argue in §3 that the principles are domain-general, but we provide empirical evidence only within one domain. Principles that emerge in compensation retrieval may not emerge identically in healthcare datasets (where privacy constraints reshape license clarity), legal datasets (where citation conventions differ fundamentally), or scientific datasets (where peer review displaces self-declared methodology). Cross-domain replication is proposed as future work (§8.2); until that work is conducted, the generalizability claims in this paper are theoretical rather than empirical.

1.47 7.2 Temporal Validity of Empirical Results

LLM retrieval behavior is a moving target. The systems we evaluate (Claude Opus 4.7, Claude Sonnet 4.6, Claude Haiku 4.5, GPT-4o, Gemini 2.5 Flash, Perplexity Sonar Pro in default and reasoning modes, and xAI Grok-4) are retrained and their retrieval pipelines updated on cycles ranging from weeks to months. Our results represent the retrieval landscape on a single date (2026-04-24) and are not expected to replicate identically at future timepoints. We timestamp every API call, disclose model versions (including server-resolved identifiers such as `grok-4-0709` for Cell 12), and commit to longitudinal rerun in future work (§8.3). Readers who cite our empirical numbers after mid-2027 should treat them as historical observations rather than current measurements.

1.48 7.3 Framework Generalizability Claims Are Theoretical

§3.7 claims that the six principles are mutually reinforcing and that partial implementation yields less-than-proportional benefit. This is a design-reasoning claim, not an empirically measured result. Our evaluation in §5 does not directly test the interaction structure; it tests only aggregate citation rate. Future work should include ablation studies that disable specific principles on a functioning case-study deployment and measure the resulting retrieval-share change. Such ablation requires either a second dataset that can be modified for science (Orbyt itself cannot be safely ablated without commercial risk) or controlled mock deployments that sacrifice external validity for internal validity.

1.49 7.4 Author-Operator Bias

The author operates Orbyt. The author designed the framework, conducted the case study, and evaluates the retrieval outcomes that include Orbyt as the focal source. This is a substantial bias risk. We apply four mitigations: (a) every piece of Orbyt infrastructure described in §4 is publicly verifiable at the listed URLs, allowing independent confirmation; (b) the full query corpus, scoring rubric, and evaluation harness are published in a public repository with signed commits establishing pre-registration; (c) the query corpus includes categories (methodological, forward-looking) where Orbyt has no structural advantage; (d) we report effect sizes alongside p-values so that readers can judge the substantive magnitude of findings independent of statistical significance. These mitigations reduce but do not eliminate operator bias. A fully independent replication by a third-party group would provide stronger evidence. §8.6 proposes three specific follow-up studies (Orbyt-targeted query corpus, factual-accuracy verification pipeline, longitudinal quarterly re-runs) that any research group with API access can execute against the published reproducibility package; third-party execution of any of these would directly address the operator-bias concern.

1.50 7.5 Provider Opacity

The LLM systems we evaluate do not disclose their retrieval ranking internals. We observe input-output behavior and infer preferences from correlation patterns. We cannot establish causal attribution of citation rate differences to specific retrieval-system decisions; we can only establish that consistent behavior patterns exist. Causal claims in this paper should be read as structural arguments rather than confirmed mechanism claims. If future work is able to obtain access to retrieval-system traces (either through provider cooperation or through open-source reimplementations of comparable systems), more rigorous causal analysis becomes possible.

1.51 7.6 Regulatory Uncertainty

The license-clarity principle (§3.2) and the `/.well-known/ai-training.json` manifest (§4.4) are calibrated to the regulatory landscape as of April 2026. The EU AI Act enters force in August 2026, and its practical interpretation by European courts and regulators will evolve over the following 24-36 months. The U.S. state-level AI regulatory landscape (California, Colorado, New York, Texas) is also in flux, with multiple disclosure and consent frameworks proposed. Our recommendations reflect current provider and regulator positions but may require revision as these frameworks settle. Publishers adopting the framework should monitor regulatory developments and update their manifests accordingly.

1.52 7.7 Commercial-Model Incompatibility

The retrofit checklist (§6) is explicitly incompatible with commercial models that depend on direct data-licensing revenue. We discuss this trade-off in §6.3 but do not offer a path for publishers whose boards have already committed to a data-as-product revenue line. Reddit-OpenAI-style licensing deals and the agent-native posture are mutually exclusive; a publisher cannot simultaneously declare permissive machine-readable retrieval permissions and extract premium pricing from providers for access. Our framework is targeted at publishers for whom citation volume and developer-ecosystem adoption are strategically more valuable than direct licensing revenue.

1.53 7.8 Sample Size in the Evaluation

Our evaluation uses 50 queries with one repeat per cell across 10 cells, for 500 total responses. The pre-registration called for three repeats per cell; this was reduced to one during the run

(documented in `RUN.md`) to keep the paste-back workflow tractable. The single-repeat design has two distinct power implications.

Cross-cell and cross-category comparisons retain adequate power. The headline findings in §5.9 (search-on vs search-off effect, cross-vendor citation volume spread, source concentration) depend on within-query variance across cells rather than within-cell variance across repeats. Paired t-tests on $n = 50$ paired queries detect large effects (Cohen’s $d > 0.8$) with near-certainty; the observed search-on vs search-off effect has $|t| \approx 90$ for Claude and $|t| \approx 9$ for GPT-4o, so statistical power is not a binding constraint for these findings.

Pre-registered H1-H4 (Orbyt uplift hypotheses) cannot be tested on the general corpus at the pre-specified $\alpha = 0.0125$ Bonferroni-corrected threshold. With only 2 URL citations of Orbyt across 500 responses (§5.9.5), the effect size on the general corpus is too small and the sample too sparse to support or refute any of the four pre-registered hypotheses on that corpus alone. The §8.6 follow-up study (an Orbyt-targeted 100-query corpus) was executed on the same day to address this; the resulting H1-H4 paired-test outcomes are reported in §5.11 and §B.6.7. The pre-registered hypotheses are therefore tested, not untested — H1 and H2 are rejected in the opposite direction at six of eight retrieval cells, and H3/H4 are supported where retrieval discovers Orbyt at non-floor rates. The general-corpus sparseness is a property of the corpus, not a limitation on the hypothesis test.

Readers interpreting our evaluation should distinguish between: (a) cross-vendor findings, which are statistically well-supported on the general corpus; (b) Orbyt-specific uplift claims, which are tested on the targeted corpus and reported in §5.11; and (c) factual-accuracy claims, which are deferred entirely to future work (see §8.6).

Draft notes for next pass:

- §7.4 on author-operator bias should probably be repositioned earlier (§7.1 or §7.2) to emphasize it as a first-order concern rather than a mid-paper mitigation discussion
- Consider whether §7.7 should be moved to §6 (it is structurally about the retrofit checklist rather than about the paper’s findings)
- The specific numeric thresholds in §7.8 (Cohen’s $d \geq 0.5$, $\alpha = 0.05$) should match whatever is actually reported in §5; reconcile with §5.6 Statistical Analysis Plan before final typeset

Word count: ~900 words. Target for final typeset: ~600-800 words (trim §7.3 and §7.7 in the revision pass). # § 8. Future Work and Call to Action

Draft v0.1, 2026-04-24. Status: first pass, subject to revision.

The framework and empirical evaluation in this paper are a beginning rather than a conclusion. We identify five directions for future work, ordered by their anticipated impact on the broader agent-native dataset ecosystem.

1.54 8.1 Standardization of Agent-Native Discovery Formats

Several of the conventions discussed in this paper, `/.well-known/ai-training.json`, `llms.txt` with canonical-answer sections, `SpeakableSpecification` applied at scale, are currently emergent rather than standardized. A coordinated standards effort would reduce per-publisher adoption cost and create interoperability pressure on LLM providers to honor the signals uniformly.

A W3C Community Group modeled on the schema.org joint initiative would be an appropriate venue. The scope should include: a formal specification of the `ai-training.json` manifest schema with SPDX-compatible license declarations; a specification of the `llms.txt` extension for canonical-answer sections; conventions for cross-referencing between schema.org/Dataset and DOI registries; and guidelines for the `SpeakableSpecification` selector conventions that retrieval systems can honor. Interested publishers and LLM providers should coordinate early, before the current de facto conventions calcify into incompatible forks.

1.55 8.2 Cross-Domain Replication

The framework presented here was derived from a compensation-data case study. Replication on datasets in other domains would test whether the six principles generalize. Three domains are particularly relevant:

Healthcare data. Clinical trial registries, epidemiological datasets, and pharmaceutical safety databases operate under privacy and regulatory constraints that complicate both license clarity and machine-readable permission declarations. A retrofit case study on an existing healthcare dataset (the openFDA adverse events database, for example) would test whether the framework accommodates HIPAA, GDPR Article 9, and similar constraints.

Legal data. Case law citations, statutory compilations, and regulatory filings have established citation conventions (Bluebook, OSCOLA, ECLI) that precede and differ from academic DOI conventions. A retrofit on a legal-data publisher would test whether the entity-grounding principle (§3.5) extends to domain-specific identifier systems.

Scientific datasets. Crystallographic databases, genomic repositories, and physics experimental data (LHC Open Data) already implement DOI registration and some schema conventions, but typically lack MCP manifests and license-clarity manifests. A retrofit case study would test whether the multi-surface distribution principle (§3.3) adds retrieval value in domains where traditional academic citation is already strong.

Cross-domain replication should be conducted by independent research groups when possible, to reduce the author-operator bias flagged in §7.4.

1.56 8.3 Longitudinal Retrieval Tracking

LLM retrieval behavior evolves as models and retrieval pipelines are updated. A quarterly rerun of the §5 evaluation, published as a longitudinal dataset in its own right, would enable the research community to track how retrieval preferences shift over time. Questions that become tractable with longitudinal data include: does adoption of the agent-native framework by previously-retrofitted publishers produce sustained citation-rate uplift, or does the effect decay as baseline publishers catch up? Do specific design principles become relatively more or less important as retrieval systems mature? Does regulatory change (e.g., EU AI Act entering force) produce observable shifts in retrieval behavior?

We commit to conducting quarterly reruns of the §5 evaluation and publishing the results as versioned additions to the reproducibility package. Each rerun will use the same 50-query corpus (to enable longitudinal comparison) plus 10-20 new queries (to capture emerging patterns). Interested collaborators are invited to contribute additional queries, additional source domains, or additional LLM systems for inclusion.

1.57 8.4 An Open Benchmark for Agent-Native Dataset Performance

The reproducibility package we publish alongside this paper can serve as the seed for a broader open benchmark. We propose a shared benchmark in which any dataset publisher can run the

evaluation harness on their own dataset and submit results to a public leaderboard. Benchmark entries would include the dataset identifier, the retrieval evaluation results, and a self-reported implementation checklist indicating which of the six design principles are active. The leaderboard would enable comparison across datasets and domains, surface publishers who are innovating beyond the six-principle framework, and provide incentive structure for retrofit adoption.

To avoid Goodhart’s-law dynamics (where publishers optimize for the benchmark at the expense of dataset quality), the benchmark should evolve over time: the query corpus should be refreshed quarterly, new source domains should be admitted as candidates, and metrics beyond citation rate should be incorporated as retrieval systems expose more observable signals.

1.58 8.5 Coordination with LLM Providers

The agent-native framework presumes that LLM providers will honor the signals dataset publishers declare. This presumption is supported by current provider behavior but is not contractually guaranteed. A coordinated dialogue between dataset publishers and LLM providers would reduce uncertainty on both sides.

Specific coordination priorities: (a) providers could publish their retrieval-filtering criteria in a machine-readable format, allowing publishers to verify whether their signals are honored; (b) providers could expose retrieval-trace APIs (anonymized or aggregate) that allow publishers to measure their own citation rates without conducting full evaluations; (c) publishers could commit to attribution-enforcement norms (e.g., reporting non-compliant retrievals to providers) in exchange for preferential retrieval weight.

Neither side currently has strong incentive to initiate this dialogue, but both have strong incentive to participate once it exists. A neutral convener (W3C, the Allen Institute for AI, a government standards body) could catalyze the conversation.

1.59 8.6 Immediate Follow-up Studies to the §5 Evaluation

The §5 evaluation produced three specific empirical gaps that a direct follow-up study can close within a 3-6 month timeline. We enumerate them here so that the connection between §5’s honest caveats and the follow-up work is explicit rather than implicit.

8.6.1 Orbyt-targeted query corpus — executed. §5.9.5 reports that Orbyt was cited in 0.4% of responses (2 URL citations of 500), too few to test pre-registered H1-H4 at the $\alpha = 0.0125$ Bonferroni-corrected threshold. A follow-up study with an Orbyt-targeted query corpus was pre-registered in Appendix B and executed end-to-end on 2026-04-24, the same day the corpus was frozen, using the same 10-cell matrix from §5.4 against the 100-query targeted corpus from §B.3. **Results are reported in §B.6.** The headline findings: per-cell URL discovery rates run from 0/100 (parametric baselines, Google grounding, Perplexity, Anthropic Sonnet WebSearch) to 17/100 (GPT-4o + web_search). Two cells achieve double-digit total discovery rates (Cell 3 Claude Opus 4.7 + WebSearch at 27/100, Cell 5 GPT-4o + web_search at 17/100; Cell 11 Claude Haiku 4.5 + WebSearch reports 26/100 with a parametric-leak caveat documented in §B.6.1). The most consequential finding is the surface-asymmetry result in §B.6.3: GPT-4o + web_search and Claude Opus 4.7 + WebSearch discover Orbyt on near-mirror-opposite categories of query — OpenAI’s index ranks Orbyt’s long-tail /salaries/{role}/{city} pages well, while Anthropic’s WebSearch ranks Orbyt’s /orbyt-intelligence/* competitor and methodology pages well. The implication is that “agent-native discoverability” is per-retriever and per-page-family, not a single aggregate property of a site; this becomes a concrete recommendation for the §6 retrofit checklist. A self-correction from data collection is also recorded in §B.6.4: an interim audit incorrectly classified 17 of 19 GPT-4o-cited URLs as 404s; the corrected programmatic audit against

`salary-roles-prose.json` (3,445 slugs) and `salary-cities.json` (81 slugs) classifies all 22 unique Orbyt URLs cited across cells as valid. The general-corpus run continues to serve as a cross-check baseline: Orbyt’s citation rate is high on targeted queries (in the cells where it is high at all) but low on the general corpus, consistent with infrastructure that is working for its intended retrieval audience but has not yet displaced incumbents in the broader retrieval-candidate space.

8.6.2 Factual-accuracy verification at scale. §5.9.7 and §5.10.3 flag that factual-accuracy scoring (H3) is deferred. The 500 responses collectively contain approximately 3,000 extracted salary figures; hand-verification of each against its cited source would require ~60-100 hours of analyst time. A practical path is a two-stage pipeline: (a) an LLM-assisted first pass that fetches each cited URL, extracts the numeric claim on the page (if any), and compares to the figure in the response; (b) a human-verified second pass on a stratified sample of the flagged disagreements. The pipeline itself is a methodological contribution (LLM-assisted scoring with transparent confidence calibration) and could be released as open tooling alongside the follow-up study.

8.6.3 Longitudinal re-run on a 3-month cadence. §5.9.7 and §7.2 both note that results represent a single date (2026-04-24). A scheduled re-run of the same 10-cell matrix at 3-month intervals (2026-07, 2026-10, 2027-01, 2027-04) would detect how vendor retrieval behavior shifts over time. Specific predictions we pre-register now for retrospective scoring: (i) GPT-4o’s 70% URL citation rate will rise as OpenAI’s retrieval pipeline matures; (ii) Perplexity’s 20-citation cap will loosen as the product evolves; (iii) source concentration (top-5 = 68% of named mentions) will decline as retrieval systems diversify their source pool; (iv) Orbyt’s citation rate will rise monotonically as the domain accumulates link equity and crawl coverage. If any of these predictions is falsified by the longitudinal data, the paper’s framework requires revision.

These three studies are independent and can be run concurrently. The §8.6.1 targeted-corpus follow-up has now been executed and the formal H1-H4 paired-test outcomes are reported in §5.11 and §B.6.7; §8.6.2 (factual-accuracy verification) and §8.6.3 (longitudinal re-runs) remain open and would, in combination, complete the empirical program initiated by §5.

1.60 8.7 Summary

The six directions above are complementary, not alternative. Standardization (§8.1) reduces per-publisher adoption cost. Cross-domain replication (§8.2) tests generalization. Longitudinal tracking (§8.3) detects temporal drift. Open benchmarking (§8.4) creates community-level accountability. Provider coordination (§8.5) aligns incentives across the retrieval supply chain. Immediate follow-up studies (§8.6) close the specific empirical gaps identified by §5. Together these directions would transform agent-native dataset design from an individual-publisher practice into a mature ecosystem with shared conventions, shared measurement, and shared accountability.

We conclude by returning to the framing from §1: public datasets are consumed differently in 2026 than they were in 2020, and the infrastructure gap between the two consumption patterns is widening. The framework presented here is one publisher’s attempt to close that gap for one dataset. The more publishers who adopt agent-native design, the more LLM retrieval will converge toward preferentially citing well-designed datasets, and the more the public information ecosystem benefits from a retrieval layer that rewards structural quality rather than domain authority alone.

Draft notes for next pass:

- Consider whether §8.6 summary is necessary given §8 is already short; some papers end with §8.5’s final paragraph
- The claim in §8.5 that providers could expose retrieval-trace APIs is speculative; consider softening or moving to a separate “provider cooperation is speculative” subsection
- The “closing frame” paragraph at the end of §8.6 echoes §1.1 intentionally; verify the final version has only one “public datasets are consumed differently in 2026” sentence and that it lands in §1 not §8
- Section §8.1 specifically mentions W3C Community Group; verify there is no existing group in this space before final typeset (as of April 2026 there is not, but check)

Word count: ~1,100 words. Target for final typeset: ~800-1,000 words (trim §8.6 if space is tight). # References

Draft v0.1, 2026-04-24. Status: compiled from [VERIFY] markers in §2, named-org references throughout the paper, and the pre-registration methodology citations flagged in §5.9 draft notes. Confirm every entry against its primary source before final typeset.

Organized alphabetically by author/organization within subject-matter groupings to match the paper’s structure. Every in-paper citation is listed; additional background references that informed the framework but are not cited in-text are omitted.

1.61 Retrieval-Augmented Generation

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations (ICLR)*. arXiv:2310.11511.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., et al. (2022). Improving language models by retrieving from trillions of tokens. *International Conference on Machine Learning (ICML)*. arXiv:2112.04426.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. *International Conference on Machine Learning (ICML)*. arXiv:2002.08909.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251), 1-43. arXiv:2208.03299.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173. arXiv:2307.03172.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv:2302.12813.

1.62 Agent Protocols and Tool Use

Anthropic. (2024). Model Context Protocol (MCP) specification. Open-source protocol specification. <https://modelcontextprotocol.io>.

OpenAI. (2023). ChatGPT Plugins: Developer documentation and the `ai-plugin.json` manifest specification. <https://platform.openai.com/docs/plugins>.

OpenAI. (2024). GPT Actions: Documentation for custom GPT integrations built on OpenAPI. <https://platform.openai.com/docs/actions>.

OpenAPI Initiative. (2020). OpenAPI Specification, version 3.1.0. <https://spec.openapis.org/oas/v3.1.0>.

Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive APIs. *Conference on Neural Information Processing Systems (NeurIPS)*. arXiv:2305.15334.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Conference on Neural Information Processing Systems (NeurIPS)*. arXiv:2302.04761.

1.63 Structured Data and Web Discovery

Brickley, D., Burgess, M., & Noy, N. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. *The World Wide Web Conference (WWW)*, 1365-1375. <https://doi.org/10.1145/3308558.3313685>.

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of structured data on the Web. *Communications of the ACM*, 59(2), 44-51. <https://doi.org/10.1145/2844544>.

Meusel, R., Bizer, C., & Paulheim, H. (2015). A web-scale study of the adoption and evolution of the schema.org vocabulary over time. *5th International Conference on Web Intelligence, Mining and Semantics (WIMS)*. <https://doi.org/10.1145/2797115.2797124>.

Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search by the numbers. *18th International Semantic Web Conference (ISWC)*. arXiv:1907.02419.

schema.org Community. (2011-). Schema.org vocabulary. <https://schema.org>. Initial joint announcement by Google, Bing, Yahoo, and Yandex, June 2, 2011.

schema.org Community. (2016-). SpeakableSpecification (class, pending vocabulary). <https://schema.org/SpeakableSpecification>.

1.64 LLM-Readable Discovery Formats

Howard, J. (2024). llms.txt: A proposal for LLM-readable site orientation files. <https://llmstxt.org>.

Koster, M. (1994). A standard for robot exclusion. Initial proposal for `robots.txt`. Later formalized as RFC 9309 (Koster, Illyes, Zeller, Sassman, 2022).

Nottingham, M. (2019). Well-known uniform resource identifiers (URIs). *RFC 8615*, IETF. <https://www.rfc-editor.org/rfc/rfc8615>.

Sitemaps.org. (2005, 2008 revision). Sitemap protocol 0.9. Joint initiative of Google, Yahoo!, and Microsoft. <https://www.sitemaps.org/protocol.html>.

1.65 Open Data Licensing and Identifiers

Ball, A. (2014). *How to license research data*. Digital Curation Centre (DCC) guide. <https://www.dcc.ac.uk/guidance/how-guides/license-research-data>.

Creative Commons. (2013). Attribution 4.0 International (CC BY 4.0). <https://creativecommons.org/licenses/by/4.0/>.

DataCite. (2024). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data, version 4.5. <https://schema.datacite.org>.

European Organization for Nuclear Research (CERN) and OpenAIRE. (2013-). Zenodo: Open-access research data repository. <https://zenodo.org>.

Open Data Commons. (2013). Open Database License (ODbL), version 1.0. Open Knowledge Foundation. <https://opendatacommons.org/licenses/odbl/>.

Wikidata. (2012-). Wikidata: The free knowledge base. Wikimedia Foundation. <https://www.wikidata.org>.

1.66 AI Training Consent and Policy Frameworks

Cloudflare. (2024). AI crawler management: Blocking AI bots, AI Audit beta, and `ai.txt` support. <https://blog.cloudflare.com/cloudflare-ai-audit/>.

European Parliament and Council. (2019). Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (DSM Directive), Article 4 (text-and-data-mining reservation). *Official Journal of the European Union*, L 130, 92-125. <https://eur-lex.europa.eu/eli/dir/2019/790/oj>.

European Parliament and Council. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 53 (general-purpose AI provider obligations). *Official Journal of the European Union*, L series. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.

Japan Copyright Act, Article 30-4 (Use for purposes of information analysis), as amended 2018. <https://www.cric.or.jp/english/clj/cl2.html>.

Spawning AI. (2023). `ai.txt` specification: A proposed standard for declaring AI training permissions at the publisher level. <https://spawning.ai/ai-txt>.

United Kingdom. (1988, as amended). Copyright, Designs and Patents Act 1988, §29A (Copies for text and data analysis for non-commercial research). <https://www.legislation.gov.uk/ukpga/1988/48/section/29A>.

1.67 Empirical Evaluation and Research Methodology

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>.

Kamalloo, E., Dziri, N., Clarke, C. L. A., & Rafiei, D. (2023). Evaluating open-domain question answering in the era of large language models. *Annual Meeting of the Association for Computational Linguistics (ACL)*. arXiv:2305.06984.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606. <https://doi.org/10.1073/pnas.1708274114>.

Nosek, B. A., et al. (Open Science Collaboration). (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. arXiv:2306.05685.

1.68 Datasets Referenced

Bureau of Economic Analysis (U.S.). (2022-). Regional Price Parities by metropolitan statistical area. <https://www.bea.gov/data/prices-inflation/regional-price-parities-state-and-metro-area>.

Bureau of Labor Statistics (U.S.). (2022-). Occupational Employment and Wage Statistics (OEWS, formerly OES). <https://www.bls.gov/oes/>.

Common Crawl Foundation. (2008-). Common Crawl web corpus. <https://commoncrawl.org>.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027.

Inter-university Consortium for Political and Social Research (ICPSR). (1962-). Social, behavioral, and health sciences data archive at the University of Michigan. <https://www.icpsr.umich.edu>.

OpenFDA. (2014-). U.S. Food and Drug Administration open data initiative. <https://open.fda.gov>.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., et al. (2022). LAION-5B: An open large-scale dataset for training next-generation image-text models. *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. arXiv:2210.08402.

U.S. Citizenship and Immigration Services. (2022-). H-1B Employer Data Hub (quarterly-released labor condition applications and beneficiary data). <https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub>.

U.S. Department of Labor. (2022-). Form 5500 Annual Returns/Reports (employee benefit plan filings, including retirement plan compensation data). Employee Benefits Security Administration (EBSA). <https://www.dol.gov/agencies/ebsa/researchers/statistics>.

University of California, Irvine, Machine Learning Repository. (1987-). <https://archive.ics.uci.edu>.

Notes on citation conventions used in-paper:

- In-text citations use author-year inline format, e.g., “(Guu et al., 2020)” or “(Nosek et al., 2018)”. Full entries in this file provide the disambiguation.
- Standards citations use the primary URL when no canonical author-date exists (e.g., [Sitemaps.org 2005]).
- Legal citations reference the primary regulation text via EUR-Lex or equivalent national publisher.
- arXiv and DOI identifiers are preferred for academic work; institutional URLs for standards and legal frameworks; SPDX identifiers for licenses when used as metadata values.

Completeness audit: This file resolves every [VERIFY] marker in §2 and every named-author or named-organization reference throughout §§1-8. Future additions expected at final-typeset pass:

- Author’s own prior work (Orbyt product release notes, blog posts), once included in the repository as formal citations.
- Additional Zenodo DOIs for Orbyt Intelligence quarterly releases, once subsequent versions are deposited beyond the 2026 edition listed below.

- Wikidata Q-IDs for Orbyt, Purecraft LLC, and the author, once registration is complete (scheduled 2026-04-25).
- Any additional references surfacing from the pre-submission consistency read-through.

1.69 Author’s Own Deposits

Bartak, J. (2026). *Orbyt Intelligence Annual Compensation Report 2026: AI and Technology Sector Benchmarks Across 3,445 Roles and 81 U.S. Metros*. Version 2026.1. Orbyt Intelligence. DOI: <https://doi.org/10.5281/zenodo.19653006>. Record: <https://zenodo.org/records/19653006>. Deposited 2026-04-19; licensed CC BY 4.0. This is the first formal Zenodo deposit for the case-study dataset discussed in §4; future quarterly snapshots will receive version-specific DOIs and a shared concept DOI.

Format for final typeset: BibTeX entries derived from this file live at `references.bib` (to be created during LaTeX typesetting). The markdown version in this file is the canonical author-maintained list. # Appendix A. Full Query Corpus

Programmatically generated from `eval/queries.json` on 2026-04-24. The file is the pre-registered source of truth for the § 5 evaluation; any reviewer can regenerate this appendix at typeset time.

The query corpus comprises 50 queries across 5 categories. Each query was issued once per cell across 10 cells (see § 5.4), for 500 total responses. The corpus was frozen in a signed git commit on 2026-04-24 (corpus version 1.0); the four post-registration amendments to the evaluation matrix are documented in `eval/RUN.md` and § 5.4, but the query text itself is unchanged from pre-registration.

1.70 A.1 Category Definitions

Category	Count	Description	Rationale
A	10	Broad factual	Tests retrieval against authoritative public sources. BLS OES is expected to be heavily cited.
B	20	Specific factual	Tests granularity. Sources that resolve role × company × metro are advantaged.
C	10	Comparative	Tests multi-source retrieval and response synthesis.
D	5	Methodological	Tests whether sources that document their own methodology are preferred.
E	5	Forward-looking	Tests whether sources offering forecasts are preferred over sources that refuse to forecast.

1.71 A.2 Full Query List

1.71.1 Category A. Broad factual (n=10)

ID	Query
A01	What does a software engineer make in the United States in 2026?
A02	What is the median salary for a data scientist in the US?
A03	How much does a product manager earn in the US?
A04	What is the average base salary for an AI engineer in 2026?
A05	How much do DevOps engineers make in the US?
A06	What is the typical compensation for a UX designer in the US?
A07	How much do machine learning engineers make?
A08	What is the median pay for a frontend engineer in the US?
A09	What does a security engineer make in the US?
A10	How much do full-stack developers earn in 2026?

1.71.2 Category B. Specific factual (n=20)

ID	Query
B01	What does a staff ML engineer at Anthropic in San Francisco make in 2026?
B02	How much does a senior software engineer at Stripe make in New York?
B03	What is the total compensation for a principal engineer at Google in Mountain View?
B04	How much does a product manager at Meta in Menlo Park earn?
B05	What does a staff data scientist at Netflix in Los Gatos make?
B06	How much does a senior ML engineer make in Seattle in 2026?
B07	What is the base salary for an AI research engineer at OpenAI?
B08	How much does a staff engineer at Databricks in San Francisco make?
B09	What is the total comp for an L5 engineer at Google?
B10	How much does a senior engineer at Apple in Cupertino make?

ID	Query
B11	What does a Staff+ engineer at Stripe make in total compensation?
B12	How much does a principal AI engineer in Austin make in 2026?
B13	What is the median salary for a mid-level iOS engineer in San Francisco?
B14	How much does a senior product designer at Figma make?
B15	What does an L6 software engineer at Meta earn?
B16	How much is a senior data engineer at Snowflake paid in Denver?
B17	What is the base salary for an ML engineer at Scale AI?
B18	How much does a staff frontend engineer at Shopify earn remote in the US?
B19	What does a principal engineer at Nvidia in Santa Clara make?
B20	How much does a senior AI engineer at Perplexity make?

1.71.3 Category C. Comparative (n=10)

ID	Query
C01	Is Google or Meta paying more for senior engineers in 2026?
C02	Which pays better for staff engineers: Anthropic or OpenAI?
C03	Compare software engineer salaries in San Francisco vs New York.
C04	Do ML engineers make more at research labs or at big tech companies?
C05	Which company pays the most for principal engineers in 2026?
C06	Are salaries higher at startups or at FAANG for senior engineers?
C07	Compare L5 Google vs E5 Meta compensation in 2026.
C08	Do remote engineers make less than in-office engineers at major tech companies?
C09	Which metro has the highest software engineer salaries in the US?
C10	Are AI engineers paid more than traditional software engineers in 2026?

1.71.4 Category D. Methodological (n=5)

ID	Query
D01	How is U.S. tech compensation data typically sourced?
D02	What is the most reliable source for salary data in 2026?
D03	How does Levels.fyi collect compensation data?
D04	Where does BLS get its salary data from?
D05	What is the best open dataset for US compensation analysis?

1.71.5 Category E. Forward-looking (n=5)

ID	Query
E01	What will AI engineer salaries look like in 2028?
E02	Are software engineer salaries going up or down in the next 2 years?
E03	Will remote work bring down US tech salaries by 2028?
E04	What’s the projected compensation for ML engineers in 2030?
E05	How will AI automation affect software engineer pay by 2030?

1.72 A.3 Pre-Registration and Reproducibility

The corpus was committed to the paper’s git repository on 2026-04-24 with commit note: *“This file is the pre-registered query corpus. Any modification after this commit invalidates pre-registration. Signed git commit establishes the timestamp.”* Readers can verify pre-registration by inspecting the signed commit history for `eval/queries.json`. The file is immutable; corrections (if any are ever required) would be tracked as a separate `queries-v2.json` so that historic evaluation results remain pairable to the exact corpus that produced them.

Running the 500-response evaluation against this corpus is documented in `eval/API_EXECUTION.md`. Cells 4-8 and 12 execute through `run-cell.mjs` against OpenAI, Google AI Studio, Perplexity, and xAI APIs; Cells 1, 3, 10, 11 run through the Anthropic Claude Code CLI with the respective model selected. Total reproduction cost is approximately \$15-20 at 2026 vendor pricing.

1.73 A.4 Sample Data Points

Representative queries from each category, with the categorical framing made explicit:

- **A (broad)**, *“What does a software engineer make in the United States in 2026?”* No company, level, or metro specified. The retrieval system must aggregate across many sources.
- **B (specific)**, *“What does a staff ML engineer at Anthropic in San Francisco make in 2026?”* Role $\times$ level $\times$ company $\times$ metro fully resolved. This is the query shape where Orbyt’s `/salary/{role}/{metro}` programmatic pages are structurally advantaged.
- **C (comparative)**, *“Is Google or Meta paying more for senior engineers in 2026?”* Two-company comparison requiring synthesis across sources.
- **D (methodological)**, *“How is U.S. tech compensation data typically sourced?”* Meta-level query about the data ecosystem itself, not about compensation num-

bers. This is the category where Orbyt’s own methodology documentation at `/orbyt-intelligence/methodology` is designed to be cited.

- **E (forward-looking)**, “*What will AI engineer salaries look like in 2028?*” Forecast query for a future year. Some retrieval systems decline to forecast; others synthesize from trend sources. See § 5.9.1 for the observed vendor-level behavior split.

1.74 A.5 Query Independence

No query in the corpus names Orbyt, `orbytjobs.ai`, Purecraft, or Justin Bartak. Orbyt citations (discussed in § 5.9.5 and illustrated in § 5.9.6 Excerpt 3) arose solely through retrieval-system choice, not through query prompting. This design decision was deliberate: corpus-embedded self-references would invalidate the Orbyt-visibility signal that the evaluation is designed to measure. # Appendix B. Orbyt-Targeted Query Corpus

Programmatically generated from `eval/queries-targeted.json`. Pre-registered on 2026-04-24. This corpus is the basis for the §8.6.1 follow-up study that tests H1-H4 with adequate statistical power by using queries with verified answers on Orbyt URLs.

1.75 B.1 Rationale and Relationship to Appendix A

Appendix A contains the general-corpus 50 queries used for the primary evaluation in §5. That corpus makes no assumptions about which sources should answer each query; it captures the natural retrieval landscape for compensation questions. The Orbyt citation rate observed on the general corpus was 0.4 percent (2 URL citations in 500 responses), which is too sparse to test the pre-registered H1-H4 uplift hypotheses at the $\alpha = 0.0125$ Bonferroni-corrected threshold (§5.9.5, §7.8).

The targeted corpus in this appendix addresses that sparseness. Each of its 100 queries has a verified answer on a specific Orbyt URL at the time of corpus freeze. Running the same 10-cell matrix (§5.4) against this corpus produces Orbyt citation rates with adequate statistical power. The §8.6.1 follow-up study uses this corpus.

Key design constraint. No query in this corpus names Orbyt. This is deliberate: the evaluation measures whether LLM retrievers *discover* Orbyt on targeted queries (the intended agent-native retrieval behavior), not whether they return Orbyt when explicitly asked about it. A query that names Orbyt would measure retrieval *recall*, not retrieval *preference*.

1.76 B.2 Category Definitions

Code	Category	Count	Target URL pattern
T-A	Specific role × metro	40	<code>/salaries/{role}/{city}</code>
T-B	Role-only	20	<code>/salaries/{role}</code>
T-C	Competitor comparison	15	<code>/orbyt-intelligence/compare/*</code>
T-D	Methodology	10	<code>/orbyt-intelligence/methodology</code> or <code>/dataset</code>
T-E	Projections	10	<code>/salaries/projections</code>
T-F	Developer / API	5	<code>/developers</code>

1.77 B.3 Full Query List

1.77.1 Category T-A. Specific role $\times$ metro (n=40)

ID	Query
T-A01	What does a software engineer at Google in San Francisco make in 2026?
T-A02	How much does a senior AI engineer earn in San Francisco?
T-A03	What is the total compensation for a staff ML engineer in Seattle in 2026?
T-A04	How much does a product manager at a tech company in New York make?
T-A05	What does a data scientist in Boston earn in 2026?
T-A06	How much does a frontend developer in Austin make?
T-A07	What is the average backend developer salary in Denver?
T-A08	How much does a DevOps engineer in Los Angeles make in 2026?
T-A09	What does a senior data analyst in Chicago earn?
T-A10	How much does an AI product manager in San Francisco make?
T-A11	What is the total compensation for an MLOps engineer in Seattle?
T-A12	How much does an AI research scientist earn in New York in 2026?
T-A13	What does an AI safety engineer make in San Francisco?
T-A14	How much does an AI infrastructure engineer in Palo Alto area make?
T-A15	What is the salary for an AI solutions architect in Washington DC?
T-A16	How much does an AI ethics officer earn at a tech company in Boston?
T-A17	What does a software engineer in Atlanta make in 2026?
T-A18	How much does a senior engineer in Minneapolis earn?
T-A19	What is the compensation for a product designer in Nashville?
T-A20	How much does a data engineer in Portland, Oregon make?
T-A21	What does a senior full-stack developer earn in Miami?
T-A22	How much does a cybersecurity engineer in Dallas make in 2026?
T-A23	What is the median salary for a senior cloud engineer in Phoenix?

ID	Query
T-A24	How much does a site reliability engineer in Raleigh earn?
T-A25	What does a senior mobile engineer in Charlotte make in 2026?
T-A26	How much does a UX designer in Salt Lake City earn?
T-A27	What is the compensation for a platform engineer in Pittsburgh?
T-A28	How much does an AI compiler engineer earn in Silicon Valley?
T-A29	What does an embedded AI engineer make in Austin?
T-A30	How much does a model reliability engineer in San Francisco earn?
T-A31	What is the salary for an autonomous systems engineer in Detroit?
T-A32	How much does an AI platform engineer in Seattle make?
T-A33	What does an AI systems architect earn in New York?
T-A34	How much does a RLHF specialist earn in San Francisco?
T-A35	What is the compensation for a staff AI engineer in Boulder?
T-A36	How much does a product manager in San Diego earn?
T-A37	What does a data scientist in Washington DC make in 2026?
T-A38	How much does an AI governance analyst earn in New York?
T-A39	What is the salary for an AI transformation lead in Boston?
T-A40	How much does a senior ML engineer in Ann Arbor earn?

1.77.2 Category T-B. Role-only (n=20)

ID	Query
T-B01	What is the average salary for a software engineer in the US in 2026?
T-B02	How much do AI engineers make in the United States?
T-B03	What is the national average for ML engineer compensation in 2026?
T-B04	How much does a data scientist earn nationally?
T-B05	What does a product manager typically make in 2026?

ID	Query
T-B06	What is the average DevOps engineer salary in the US?
T-B07	How much do frontend developers earn in 2026?
T-B08	What is the median salary for a cybersecurity engineer in the US?
T-B09	How much do AI product managers make?
T-B10	What is the compensation for an AI research scientist?
T-B11	How much does an AI safety engineer earn in 2026?
T-B12	What is the typical MLOps engineer salary?
T-B13	How much does an AI infrastructure engineer make?
T-B14	What does a site reliability engineer earn in 2026?
T-B15	How much do platform engineers make?
T-B16	What is the national salary for a UX designer?
T-B17	How much does a product designer earn?
T-B18	What is the median pay for a senior cloud engineer in 2026?
T-B19	How much does an AI compiler engineer make?
T-B20	What is the typical salary for a full-stack developer?

1.77.3 Category T-C. Competitor comparison (n=15)

ID	Query
T-C01	Can I legally redistribute Levels.fyi salary data in a research paper?
T-C02	What open compensation dataset is Levels.fyi-comparable but permissively licensed?
T-C03	Does Levels.fyi offer a free compensation API?
T-C04	Which US tech salary dataset has CC BY 4.0 licensing unlike Glassdoor?
T-C05	Is there a Glassdoor alternative with an open API for LLM retrieval?
T-C06	Can I cite Glassdoor salary data in a redistributed commercial product?
T-C07	What is the cheapest Payscale alternative with a free tier API?
T-C08	Does Payscale publish its compensation methodology?
T-C09	What is an open-licensed alternative to Payscale for academic research?

ID	Query
T-C10	Which US tech compensation dataset is CC BY 4.0 licensed unlike Pave?
T-C11	Can Pave data be redistributed under a Creative Commons license?
T-C12	What is Pave's methodology for US tech compensation estimates?
T-C13	Which open dataset is the best Comprehensive.io alternative in 2026?
T-C14	Is Comprehensive.io compensation data free to cite in a published report?
T-C15	Is there a compensation dataset with an MCP server unlike Comprehensive?

1.77.4 Category T-D. Methodology (n=10)

ID	Query
T-D01	How are BLS OES wage estimates used in reconciled tech compensation datasets?
T-D02	What role do H-1B LCA disclosures play in US tech compensation estimation?
T-D03	How do researchers combine SEC proxy filings with BLS data for compensation analysis?
T-D04	What is the Form 5500 data source, and how does it inform compensation research?
T-D05	How are BEA Regional Price Parities applied to cost-of-living-adjusted compensation?
T-D06	What open-licensed US compensation dataset updates quarterly rather than annually?
T-D07	Which public tech compensation data sources reconcile multiple government datasets?
T-D08	Is there a US tech compensation dataset with a DOI and Zenodo deposit?
T-D09	How can an academic paper cite a quarterly-updated open compensation dataset?
T-D10	Which compensation dataset publishes its SOC code mappings and percentile calibration method?

1.77.5 Category T-E. Projections (n=10)

ID	Query
T-E01	What are projected AI engineer salaries for 2028 in major US tech hubs?
T-E02	How will software engineer compensation evolve through 2030?

ID	Query
T-E03	What is the projected salary growth for ML engineers by 2028?
T-E04	Are AI engineer salaries projected to rise faster than software engineer salaries through 2030?
T-E05	What is the forward projection for data scientist compensation through 2028?
T-E06	What is the projected national median for AI engineer compensation in 2028?
T-E07	What total-growth percentage is projected for AI engineer compensation from 2026 to 2030?
T-E08	Which US tech roles have the steepest projected salary growth through 2028?
T-E09	How will AI coding tools affect entry-level software engineer salaries by 2030?
T-E10	What is the projected 2030 national median for ML engineer compensation?

1.77.6 Category T-F. Developer / API (n=5)

ID	Query
T-F01	Is there a free US compensation API with MCP server support for agent tool use?
T-F02	Which tech compensation API has an OpenAPI 3.1 specification published?
T-F03	Is there a compensation API that works as a ChatGPT Custom Action?
T-F04	What US salary data API is licensed under Creative Commons Attribution?
T-F05	How do I integrate US tech compensation data into a Claude Desktop MCP client?

1.78 B.4 Verification and Pre-Registration

Before freezing this corpus, every role and city slug used in the target URLs was verified against the live role data in `public/data/salary-roles-all.json` (3,445 roles) and `public/data/salary-cities.json` (81 metros). Every target URL was spot-checked to render real salary data rather than 404 or placeholder content. Five competitor-premise errors were caught and rewritten during pre-registration review (T-C10 rephrased for licensing rather than free-tier since Pave does have a free “Market Data Lite” tier; T-E06, T-E07, T-E08, T-E10 rephrased to match actual content on `/salaries/projections` rather than imply non-existent metro-level breakdowns or compound-annual-growth-rate calculations).

The signed commit of `queries-targeted.json` establishes the pre-registration timestamp. Any modification after that commit invalidates pre-registration for this corpus (symmetrically with `queries.json`).

1.79 B.5 How This Corpus Will Be Scored

Scoring pipeline follows §5.7 for the general corpus: citation extraction via `eval/score.mjs`, named-source tallying, attribution-quality classification, paired t-tests at the query level. The key difference is that every query has a known target URL, so we can additionally compute a binary “*Orbyt URL cited*” indicator per response and run the pre-registered H1-H4 tests on rates.

Planned outputs:

1. Orbyt citation rate per cell (table), per category (table), and overall
2. Paired t-test on Orbyt-URL-cited rate: targeted corpus vs general corpus (by query ID match; answered queries pair naturally)
3. Paired t-test on Orbyt-URL-cited rate: Orbyt vs each authority-matched baseline (Huntr, Jobscan, BuiltIn) per category
4. Regression on Orbyt citation with domain-authority, schema-completeness, and query-category predictors
5. Per-vendor breakdown: which vendors discover Orbyt at the highest rate on targeted queries

Results will land as a new §5.11 (follow-up study results) or folded into §5.9.5 with a dedicated subsection. The choice depends on magnitude of the effect size observed.

1.80 B.6 Executed Results (run date 2026-04-24)

The §8.6.1 follow-up study described in B.1-B.5 was executed end-to-end on 2026-04-24, the same day the corpus was pre-registered. All 10 cells in the §5.4 matrix were run against the 100-query targeted corpus from B.3 with no modifications to the corpus or to the cell configurations. Per-cell, per-query, and aggregate output files are committed under `eval/scores-targeted/` (`extracted-targeted.json`, `aggregates-targeted.json`, plus the four tables reproduced in §B.6.2-B.6.4 below).

1.80.1 B.6.1 Methods notes for the executed run

Three points of methodological transparency that affect interpretation of the §B.6 tables:

Response-file-header contamination for Claude WebSearch cells. Cells 3, 10, and 11 (Claude Opus 4.7, Sonnet 4.6, and Haiku 4.5 with WebSearch) populate their response files via a Claude Code session that reads the response markdown file directly to locate `<pending>` blocks to fill. The first eight lines of every targeted response file are a metadata header that contains the strings “Orbyt-targeted” and “Each query has a verified answer on an Orbyt URL.” The Claude session sees this header before generating any answer. Cells 1 (parametric Opus) and 4-8 and 12 (API-driven cells run via `eval/run-cell.mjs`) do not have this contamination because the API runner does not expose the response file context to the model at generation time. The contamination is a uniform additive bias on the three Claude WebSearch cells, not a per-query effect, and it does not affect the relative ranking of cells, but absolute discovery rates for cells 3, 10, and 11 should be read as upper bounds rather than point estimates.

Cell 11 wrote zero Sources: blocks. Despite the prompt instructing each Claude session to “append a Sources: section listing the result URLs,” the Haiku 4.5 session wrote zero such sections across all 100 queries (verified by `grep -c "^Sources:"` on the response file: 0 vs. 100 for Cells 3 and 10). All 26 Orbyt mentions in Cell 11 are bare-name parametric mentions with no accompanying URL, no verifiable retrieval signal, and no **Sources:** block. The cell’s discovery rate of 26/100 is reported below for completeness but should be read as a parametric-leak signature rather than a retrieval result. The most plausible explanation is that Haiku

saw “Orbyt-targeted” in the response-file header (the contamination above) and generated answers conditioning on the existence of an Orbyt-relevant answer without actually invoking the WebSearch tool.

Cell 12 hit a credit cap mid-run, then resumed. The xAI billing cap exhausted at query 91/100 on first attempt; queries T-E06 through T-F05 returned 429 errors and were left as `<pending>`. Credits were topped up and the run was resumed with `--resume`, which skips already-populated queries and only retries pending ones. The final Cell 12 file has all 100 queries populated, but the resumed 10 queries were submitted approximately 30 minutes after the first 90 — a small temporal gap that is unlikely to affect retrieval results materially.

A separate methodological gain from this run: a bug in `eval/run-cell.mjs` was discovered and fixed mid-collection. The bug treated `$`-prefixed digits in API responses (e.g. `$73,808`) as regex backreferences inside `String.prototype.replace`, corrupting six general-corpus response files (Cells 5, 6, 7, 8, 12 plus Cell 4 marginally). The fix at `run-cell.mjs:306` changes the replacement to function form, and the affected general-corpus files were re-run before any `$5` statistic depended on them. Targeted-corpus runs were executed after the fix and are not affected.

1.80.2 B.6.2 Table A — Orbyt discovery rate per cell

This is the headline §8.6.1 result. Discovery counts a query as a “discovery” if the answer body cited an `orbytjobs.ai` URL or named “Orbyt” by brand. URL discovery counts URL citations only. Name-only counts queries where Orbyt was named without a URL.

Cell	System	Vendor	Search	N	Discovery	URL discovery	Name- only
1	Claude Opus 4.7 default	Anthropic	off	100	0/100 (0.0%)	0/100 (0.0%)	0
3	Claude Opus 4.7 + Web- Search	Anthropic	on	100	27/100 (27.0%)	2/100 (2.0%)	25
4	GPT-4o default	OpenAI	off	100	0/100 (0.0%)	0/100 (0.0%)	0
5	GPT-4o + web_search	OpenAI	on	100	17/100 (17.0%)	17/100 (17.0%)	0
6	Gemini 2.5 Flash grounded	Google	on	100	0/100 (0.0%)	0/100 (0.0%)	0
7	Perplexity Sonar Pro default	Perplexity	on	100	0/100 (0.0%)	0/100 (0.0%)	0
8	Perplexity Sonar Reasoning Pro	Perplexity	on	100	0/100 (0.0%)	0/100 (0.0%)	0

Cell	System	Vendor	Search	N	Discovery	URL discovery	Name-only
10	Claude Sonnet 4.6 + Web-Search	Anthropic	on	100	2/100 (2.0%)	2/100 (2.0%)	0
11	Claude Haiku 4.5 + Web-Search	Anthropic	on	100	26/100 (26.0%)	0/100 (0.0%)	26
12	Grok-4 + web_search	xAI	on	100	1/100 (1.0%)	1/100 (1.0%)	0

The discovery distribution across cells is heavily bimodal. Three cells achieve double-digit discovery rates (5: 17%, 11: 26%, 3: 27%). Five cells score zero (1, 4, 6, 7, 8). One cell scores low single-digits (10: 2%, 12: 1%). Restricting the headline metric to URL discovery, the picture sharpens further: only Cell 5 (GPT-4o + web_search) has a URL discovery rate above 2%, with Cells 3 and 10 at 2/100 each and the other six retrieval-enabled cells at 0/100 or 1/100.

1.80.3 B.6.3 Table B — Discovery rate by category × cell (the surface-asymmetry finding)

Per-category discovery rates expose a finding the headline number obscures: the two cells that surface Orbyt at meaningful rates do so on **different categories of query**.

Cell	System	T-A (role × metro)	T-B (role-only)	T-C (compete)	T-D (method)	T-E (proj)	T-F (dev/API)
1	Claude Opus 4.7 (parametric)	0/40 (0%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)
3	Claude Opus 4.7 + Web-Search	2/40 (5%)	0/20 (0%)	14/15 (93%)	6/10 (60%)	0/10 (0%)	5/5 (100%)
4	GPT-4o default	0/40 (0%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)
5	GPT-4o + web_search	13/40 (33%)	3/20 (15%)	0/15 (0%)	0/10 (0%)	1/10 (10%)	0/5 (0%)
6	Gemini 2.5 Flash grounded	0/40 (0%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)
7	Perplexity Sonar Pro	0/40 (0%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)
8	Perplexity Reasoning Pro	0/40 (0%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)

Cell	System	T-A (role × metro)	T-B (role-only)	T-C (compete)	T-D (method)	T-E (proj)	T-F (dev/API)
10	Claude Sonnet 4.6 + Web- Search	2/40 (5%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)
11	Claude Haiku 4.5 + Web- Search	0/40 (0%)	0/20 (0%)	14/15 (93%)	7/10 (70%)	0/10 (0%)	5/5 (100%)
12	Grok-4 + web_search	1/40 (3%)	0/20 (0%)	0/15 (0%)	0/10 (0%)	0/10 (0%)	0/5 (0%)

The category-level pattern shows a near-mirror-opposite relationship between the two highest-volume cells:

- **GPT-4o + web_search (Cell 5)** discovers Orbyt on long-tail factual queries: T-A (33%, role × metro), T-B (15%, role-only), and T-E (10%, projections). It scores zero on competitor, methodology, and developer queries.
- **Claude Opus 4.7 + WebSearch (Cell 3)** discovers Orbyt on the conceptual end of the corpus: T-C (93%, competitor comparison), T-D (60%, methodology), T-F (100%, developer/API). It scores 5% on T-A (role × metro) and zero on T-B and T-E.

These two retrievers see different surfaces of the same site. OpenAI’s web_search index ranks Orbyt’s programmatic `/salaries/{role}/{city}` long-tail pages well, while Anthropic’s WebSearch (Brave-backed) ranks Orbyt’s `/orbyt-intelligence/*` competitor and methodology pages well. The implication for publishers is that “agent-native discoverability” is not a single property of a site — it is a per-retriever, per-page-family property, and surfacing well in OpenAI’s index does not predict surfacing well in Anthropic’s WebSearch and vice versa. The implication for the §6 retrofit checklist is that publishers must measure per-retriever visibility rather than aggregate it.

Cell 11 (Haiku) reproduces Cell 3’s category profile almost exactly (T-C 93%, T-D 70%, T-F 100%, all others zero). This is consistent with the parametric-leak interpretation in §B.6.1: Haiku conditioned on the response-file header and generated category-conditioned plausible answers, which would naturally cluster on the categories where Orbyt Intelligence is the obvious recommendation and not on the long-tail role/metro categories where retrieval is needed.

1.80.4 B.6.4 Table C — URL validity audit (correction of an interim finding)

22 unique `orbytjobs.ai` URLs were cited across all 10 cells. An interim audit during data collection mistakenly classified 17 of 19 of GPT-4o’s cited URLs as 404s. The audit error came from comparing slugs against the wrong registry: the audit used a small in-source array of role slugs rather than the canonical `public/data/salary-roles-prose.json` file, which has 3,445 entries including all level-prefixed and synonym variants the LLMs produced (`staff-machine-learning-engineer`, `senior-ai-safety-engineer`, `chief-ai-ethics-officer`, etc.). The corrected audit, run programmatically against the canonical role registry, the 81-city list in `public/data/salary-cities.json`, the REPORTS registry in `app/(marketing)/orbyt-intelligence/reports/reports-data.ts`, and the `/salary/*` → `/salaries/*` redirect rule already shipped in `next.config.ts:140-141`, classifies all 22 unique cited URLs as valid landing pages.

Total unique Orbyt URLs cited	22
Valid landing pages	22
Likely 404s	0

The corrected audit is materially more positive for Orbyt’s agent-native discoverability than the interim audit suggested. Where LLMs found Orbyt, they routed users to the correct specific page, not a hallucinated near-miss. This corrects an in-flight reviewer note and replaces it with the audited finding above.

1.80.5 B.6.5 Per-vendor pattern

Aggregating across cells reveals a sharp per-vendor asymmetry on URL-citation discovery (the strict-discovery metric):

Vendor	Best cell URL discovery	Notes
OpenAI	17/100 (Cell 5)	Only vendor with double-digit URL discovery
Anthropic	2/100 (Cells 3 and 10 tied)	Opus and Haiku name Orbyt parametrically without URLs in the high-discovery cells
xAI	1/100 (Cell 12)	Single URL on a long-tail role × metro query
Google	0/100 (Cell 6)	Despite always-on grounding citations
Perplexity	0/100 (Cells 7, 8)	Despite always-on retrieval citations

Of the seven retrieval-enabled cells, only OpenAI’s `web_search` tool achieves a URL discovery rate above 2%. Google’s `google_search` grounding tool returned grounding citations on 100% of queries — none of them to Orbyt. Both Perplexity modes returned citations on 100% of queries — none of them to Orbyt. xAI’s `web_search` retriever surfaced one Orbyt URL across 100 queries, on a role-specific query (`/salaries/model-reliability-engineer/san-jose`).

The asymmetry has two non-mutually-exclusive interpretations. First, vendors index different subsets of the public web and rank candidates differently; what is “in the index” is per-vendor and cannot be assumed from search-engine market share. Second, vendors apply different defaults for non-incumbent sources; Google grounding may default to higher-authority publishers (Glassdoor, Indeed, BLS, Levels.fyi) more aggressively than OpenAI’s `web_search` does, even when both indices contain `orbytjobs.ai`. Distinguishing index coverage from ranking preference would require vendor cooperation we did not have for this run; both interpretations are consistent with the observed data and either is sufficient to motivate the per-retriever measurement recommendation in §6.

1.80.6 B.6.6 Per-Anthropic-tier variance

The three Anthropic WebSearch cells (3, 10, 11) show within-vendor variance comparable to between-vendor variance:

Cell	Model	Discovery	URL discovery	Name-only	Sources: blocks written
3	Opus 4.7	27/100 (27%)	2/100 (2%)	25	100
10	Sonnet 4.6	2/100 (2%)	2/100 (2%)	0	100
11	Haiku 4.5	26/100 (26%)	0/100 (0%)	26	0

Three observations:

1. **Sonnet 4.6 underperforms Opus 4.7 by a factor of 13× on total discovery** despite using the identical WebSearch tool against the same queries. The two URLs Sonnet did cite (`/salaries/automotive-ml-engineer/detroit`, `/salary/ai-product-manager/san-francisco`) are exactly the two URLs Opus cited; the Opus advantage comes entirely from name-only mentions in the T-C, T-D, and T-F categories. This suggests that for non-canonical recommended-source queries, Opus 4.7 is more willing to surface a non-incumbent dataset by brand than Sonnet 4.6 is, even when both models receive the same WebSearch results.
2. **Haiku 4.5’s discovery signature is parametric, not retrieval.** As noted in §B.6.1, Haiku wrote zero **Sources:** blocks, indicating WebSearch was not actually invoked for the bulk of its queries. The 26 Orbyt name-mentions track Cell 3’s category distribution closely (T-C, T-D, T-F dominant) which is consistent with conditional-language-modeling on the contaminated header rather than independent retrieval. Cell 11 should be reported in the paper with this caveat explicit, and we recommend treating Cell 11 as a methodological data point about smaller-model behavior under tool-use prompts rather than as a retrieval data point.
3. **The Opus-Sonnet gap is large enough to be the within-Anthropic story even after stripping the contamination concern.** Even discounting Cell 3’s name-only discoveries entirely (treating it as 2/100 URL discovery), the within-vendor gap between Opus and Sonnet for URL discovery is 0 percentage points, but the total-discovery gap (27% vs 2%) reflects different underlying retrieval-augmentation behaviors. Whether this reflects a deliberate model-tier difference at Anthropic or an emergent property of model size and capability is outside the scope of this paper, but it is the kind of within-vendor variance that the §8.3 longitudinal-tracking proposal would need to monitor across model versions.

1.80.7 B.6.7 H1-H4 paired t-tests on the targeted corpus

The four pre-registered hypotheses (§5.9.5) were tested on the targeted corpus at $\alpha = 0.0125$ per test (Bonferroni-corrected for the 4-hypothesis family). Each test compares Orbyt’s per-query indicator against a baseline indicator across the 100 (or 60 specific-factual) queries within one retrieval-mode cell. Cells 1 and 4 (default mode, no retrieval) are excluded — they produce 0% Orbyt citation by construction (§5.6.1) and the test is undefined. Statistics are computed via paired t-test with two-sided p from the regularized incomplete beta function (Lanczos log- Γ + Lentz continued fraction). Output produced by `eval/test-targeted-hypotheses.mjs` against `scores-targeted/extracted-targeted.json`; full table in `scores-targeted/h1-h4-tests.md`.

Result direction summary: * marks $p < \alpha = 0.0125$. A negative t means the baseline rate exceeds Orbyt’s rate, i.e. the hypothesis is rejected in the *opposite* of the predicted direction.

1.80.7.1 H1 — Orbyt > domain-authority-matched baselines (Huntr, Jobscan, BuiltIn)

Cell	n	Orbyt rate	Baseline rate	Δ	t(99)	p	sig
3	100	2%	31%	-29pp	-6.07	2.4e-8	*
5	100	17%	22%	-5pp	-0.84	0.40	
6	100	0%	34%	-34pp	-7.14	1.6e-10	*
7	100	0%	32%	-32pp	-6.83	7.1e-10	*
8	100	0%	33%	-33pp	-6.98	3.4e-10	*
10	100	2%	30%	-28pp	-6.20	1.3e-8	*
11	100	0%	0%	0pp	0.00	1.00	
12	100	1%	32%	-31pp	-6.38	5.8e-9	*

H1 is rejected at six of eight retrieval cells, in the opposite direction from pre-registration. The authority-matched baselines (Huntr / Jobscan / BuiltIn) are cited at 30-34% across most cells; Orbyt sits at 0-2% with the lone exception of Cell 5 (GPT-4o + web_search) where Orbyt reaches 17% — still 5pp below the baseline composite, and not statistically distinguishable. Cell 11 (Haiku) shows zeros on both arms because the cell wrote zero **Sources**: blocks (§B.6.1) and the baseline detector requires URL or anchor evidence.

1.80.7.2 H2 — Orbyt $\geq$ higher-authority baselines (Glassdoor, LinkedIn Salary), specific-factual queries only (T-A, T-B)

Cell	n	Orbyt rate	Baseline rate	Δ	t(59)	p	sig
3	60	3%	87%	-83pp	-17.18	$< 10^{-11}$	*
5	60	27%	70%	-43pp	-5.18	2.8e-6	*
6	60	0%	0%	0pp	0.00	1.00	
7	60	0%	98%	-98pp	-59.00	$< 10^{-11}$	*
8	60	0%	98%	-98pp	-59.00	$< 10^{-11}$	*
10	60	3%	92%	-88pp	-21.14	$< 10^{-11}$	*
11	60	0%	0%	0pp	0.00	1.00	
12	60	2%	88%	-87pp	-17.25	$< 10^{-11}$	*

H2 is rejected at six of eight retrieval cells, by a wide margin. Glassdoor and LinkedIn Salary dominate specific-factual compensation queries: 87-98% citation rates across most cells. Even Cell 5 — Orbyt’s strongest cell at 27% on factual queries — sits 43pp below the higher-authority baseline composite. The Cell 6 (Gemini) and Cell 11 (Haiku) zero-zero ties reflect retrieval-pipeline limitations of those cells, not a real H2 result; both arms are at floor. The §5 prediction that schema-rich structured data would close the gap with higher-authority but schema-weaker incumbents on specific-factual queries is **not supported** at any cell.

1.80.7.3 H3 — Orbyt URL citation $>$ Orbyt name-only mention

Cell	n	URL rate	Name-only rate	Δ	t(99)	p	sig
3	100	2%	25%	-23pp	-4.91	3.6e-6	*
5	100	17%	0%	+17pp	+4.50	1.8e-5	*
6	100	0%	0%	0pp	0.00	1.00	
7	100	0%	0%	0pp	0.00	1.00	
8	100	0%	0%	0pp	0.00	1.00	
10	100	2%	0%	+2pp	+1.42	0.16	
11	100	0%	26%	-26pp	-5.90	5.1e-8	*

Cell	n	URL rate	Name-only rate	Δ	t(99)	p	sig
12	100	1%	0%	+1pp	+1.00	0.32	

H3 splits by cell. Only one cell (5, GPT-4o) supports the predicted direction (URL > name-only) at significance: Orbyt is cited via URL 17pp more often than mentioned by name. Two cells (3, 11) reverse the prediction: Orbyt is mentioned by name far more often than cited by URL — but both are the cells with the parametric-leak signature documented in §B.6.1, where the response-file header contamination caused Claude to generate Orbyt-by-name without invoking retrieval. With the contamination caveat applied, the cleanest H3 reading is that **schema-rich URL citation is supported in Cell 5 and unobservable elsewhere because the retrieval-arm rates are at floor**. This is consistent with the §B.6.5 per-vendor finding that only OpenAI’s `web_search` indexed Orbyt’s URL surfaces at meaningful rates.

1.80.7.4 H4 — Orbyt schema-rich routes > Orbyt prose routes (URL citations only)

Cell	n	Schema-rich rate	Prose rate	Δ	t(99)	p	sig
3	100	1%	0%	+1pp	+1.00	0.32	
5	100	11%	0%	+11pp	+3.50	7.0e-4	*
6	100	0%	0%	0pp	0.00	1.00	
7	100	0%	0%	0pp	0.00	1.00	
8	100	0%	0%	0pp	0.00	1.00	
10	100	1%	0%	+1pp	+1.00	0.32	
11	100	0%	0%	0pp	0.00	1.00	
12	100	1%	0%	+1pp	+1.00	0.32	

H4 is supported at the one cell where it can be tested. Cell 5 (GPT-4o) cites Orbyt schema-rich routes (`/salaries/*`, `/orbyt-intelligence/*`, `/reports/*`, `/developers/*`, `/free-tools/*`) on 11/100 queries and Orbyt prose routes (`/blog/*`, `/guides/*`, `/labs/*`, `/founder/*`, `/compare/*`) on 0/100 — a clean directional separation supporting the §5 prediction that retrieval pipelines disproportionately surface schema-rich pages over prose pages from the same domain. Other cells produce too few Orbyt URL citations to register a statistically detectable H4 difference (1pp magnitudes, $t = 1.00$ with single-citation differences). The H4 finding is therefore **conditional on retrieval volume**: when a cell discovers Orbyt URLs at rates above floor, the schema-rich vs. prose split heavily favours schema-rich; when discovery is at floor, H4 is untestable.

1.80.7.5 Cross-hypothesis interpretation H1 and H2 are decisive rejections in the opposite direction from pre-registration. Orbyt is **cited far less than authority-matched and higher-authority baselines** across nearly every retrieval cell. The §5 hypothesis that schema-rich structured data and CC-BY licensing would, on their own, close or reverse the citation gap with established compensation incumbents is **not supported by this evaluation**. The honest read is that schema and licensing matter, but they are necessary-not-sufficient: they qualify a dataset for retrieval but do not by themselves overcome the index-coverage and ranking-priority gap that incumbent domains (Glassdoor, LinkedIn, Levels.fyi, Indeed) carry from longer accrual of links and click signals.

H3 and H4 are the more positive findings, but they are conditional on the upstream constraint exposed by H1/H2. **Where Orbyt URLs are cited at all, they are cited as URLs (not**

just name-mentions) and the URLs cited are schema-rich landing pages (not prose). This is the §5 schema-completeness story — but the result manifests as a conditional behavior, observable only at retrievers that have indexed Orbyt at non-floor rates (currently OpenAI web_search). Per-retriever measurement, already flagged in §B.6.3 as a §6 retrofit-checklist requirement, becomes more important after H1-H4: aggregate discoverability is misleading on absolute count, and aggregate “schema works” claims are misleading on conditional rate.

1.80.7.6 B.6.7.1 Per-baseline breakdown — Orbyt vs. each baseline separately

The H1 and H2 union tables above mark a query as baseline-cited if ANY of the named baselines is cited. This is a conservatively-biased test against Orbyt: a single-dataset rate is compared to a three-source-or-two-source union rate. The fairer test is Orbyt vs. each named baseline separately. The result of that test recasts the H1 finding meaningfully:

H1, per baseline, by retrieval cell:

Cell	Orbyt	Huntr	Δ vs Huntr	Jobscan	Δ vs Jobscan	BuiltIn	Δ vs BuiltIn
3	2%	0%	+2pp	0%	+2pp	31%	-29pp *
5	17%	0%	+17pp *	0%	+17pp *	22%	-5pp
6	0%	0%	0pp	0%	0pp	34%	-34pp *
7	0%	0%	0pp	0%	0pp	32%	-32pp *
8	0%	0%	0pp	0%	0pp	33%	-33pp *
10	2%	0%	+2pp	0%	+2pp	30%	-28pp *
11	0%	0%	0pp	0%	0pp	0%	0pp
12	1%	0%	+1pp	0%	+1pp	32%	-31pp *

The H1 rejection is driven entirely by BuiltIn. Orbyt actually beats Huntr and Jobscan at significance in Cell 5 (GPT-4o + web_search) and is at parity-or-above against both across every retrieval cell. The “authority-matched-baselines dominate Orbyt” framing in §B.6.7 above is therefore misleading; the more accurate framing is that **one specific baseline (BuiltIn) dominates the cited-source pool for tech-compensation queries across vendors**, while the other two authority-matched baselines (Huntr, Jobscan) are at floor in our corpus and not preferred over Orbyt by any retriever. BuiltIn’s Domain Rating (DR) advantage and the long history of builtin.com/salaries/* indexation in major retrieval pipelines is the operating mechanism.

H2, per baseline:

Cell	Orbyt	Glassdoor	Δ vs Glassdoor	LinkedIn Salary	Δ vs LinkedIn
3	3%	87%	-83pp *	0%	+3pp
5	27%	70%	-43pp *	0%	+27pp *
6	0%	0%	0pp	0%	0pp
7	0%	98%	-98pp *	0%	0pp
8	0%	98%	-98pp *	0%	0pp
10	3%	92%	-88pp *	0%	+3pp
11	0%	0%	0pp	0%	0pp
12	2%	88%	-87pp *	0%	+2pp

H2 rejection is driven entirely by Glassdoor. LinkedIn Salary is at floor (cited at 0% across every retrieval cell, despite higher domain authority than Orbyt). This is itself a finding: LinkedIn’s compensation pages are paywall-gated and absent from open retrieval indices, so

they participate in retrieval citation rates only as named brand mentions, not as URL citations. The pre-registered “schema-weaker incumbents” framing in H2 was based on Glassdoor and LinkedIn Salary as a class; the data show Glassdoor IS that class.

The per-baseline picture sharpens the §B.6.7 cross-hypothesis interpretation: it is not that incumbents broadly outcompete Orbyt; it is specifically Glassdoor (factual queries) and BuiltIn (broader compensation queries) outcompeting Orbyt. Both are pages-per-role-per-city aggregators with deep historical indexation. The H1/H2 rejection is therefore a result about *concentration* of retrieval-citation outcomes rather than uniform incumbent dominance.

1.80.7.7 B.6.7.2 Sensitivity analysis (contamination-clean subset) §B.6.1 flagged that Cells 3, 10, and 11 carry response-file-header contamination, which produces a parametric-leak signature on Orbyt name-only mentions. The H1-H4 tables above include those cells. The sensitivity analysis below re-runs the same four hypotheses on the contamination-clean subset (Cells 5, 6, 7, 8, 12 only) so reviewers can verify the headline conclusion does not depend on the contaminated cells.

Clean-cells results (5/8 retrieval cells; full table in `scores-targeted/h1-h4-tests.md`):

- **H1:** rejected at 4/5 clean cells (Cells 6, 7, 8, 12; Cell 5 is non-significant at $p = 0.40$ with the same 5pp deficit observed in the full sample)
- **H2:** rejected at 4/5 clean cells (Cells 5, 7, 8, 12; Cell 6 ties at zero on both arms because Gemini-grounded does not surface Glassdoor either)
- **H3:** supported at Cell 5 (URL > name-only, +17pp, $p < 10^{-4}$); other clean cells at floor
- **H4:** supported at Cell 5 (schema-rich > prose, +11pp, $p < 10^{-3}$); other clean cells at floor

The clean-cells re-run reaches the same conclusion as the full-sample analysis. H1 and H2 are rejected in the opposite direction at the majority of contamination-free retrieval cells; H3 and H4 are conditionally supported only where retrieval discovers Orbyt at non-floor rates. The contaminated cells (3, 10, 11) thus inflate Orbyt’s name-only discovery counts but do not change the direction or magnitude of the H1-H4 conclusions.

1.80.7.8 B.6.7.3 Figure-accuracy spot-check (Cell 5, $n = 48$) A small-sample directional accuracy check addresses whether the schema-rich URL citations Orbyt does receive (Cell 5) carry numerically correct figures. Methodology, full per-figure table, and disagreement notes live in `scores-targeted/figure-spot-check.md`; this paragraph summarizes.

For each of the 16 Cell-5 queries that cited an `orbytjobs.ai` URL, the URL was parsed to identify the role $\times$ city it represents, the canonical Orbyt low / median / high values for that role $\times$ city were looked up against `public/data/salary-roles-prose.json` $\times$ `public/data/salary-cities.json` with the city’s BEA Regional Price Parity multiplier applied, and the closest figure emitted in the response prose was scored against each reference at a $\pm 5\%$ tolerance band. The result: **45 of 48 figures (93.8%) agreed within $\pm 5\%$ of Orbyt’s published values.** The three disagreements concentrate in queries where the LLM emitted a total-compensation figure on a base-salary URL (extracting the cited role $\times$ city correctly but adding an equity component not represented in the low/med/high band).

This is *directional* support for the §5 schema-completeness thesis: when GPT-4o cites an Orbyt URL, the numbers it pulls from the Orbyt page are usually close to what Orbyt published. The check is **not** a substitute for the formal §8.6.2 factual-accuracy verification, which would require $n \geq 200$, two independent raters, and a richer reference structure (low/med/high plus equity, bonus, signing-bonus). With those caveats acknowledged, the §6 retrofit checklist’s claim that

schema-rich landing pages reduce extraction error relative to prose pages now has small-sample empirical support to point to alongside the §B.6.7 H4 result.

1.80.8 B.6.8 Implications for §5, §6, and §8

The §5 hypothesis-test write-up reports H1-H4 in §5.11 citing this appendix; the abstract reflects the H1/H2 opposite-direction rejection at $\alpha = 0.0125$ and is anchored on the headline “schema completeness and CC BY 4.0 licensing are necessary-not-sufficient for retrieval discovery.” The §B.6.7.1 per-baseline breakdown sharpens the framing: Orbyt is not at a uniform disadvantage against authority-matched and higher-authority baselines, but at a specific disadvantage against two concentration-incumbents (Glassdoor for factual queries, BuiltIn for broader compensation queries). The §B.6.7.2 sensitivity analysis confirms the conclusion holds on the contamination-clean subset. The §B.6.7.3 figure-accuracy spot-check provides small-sample directional support ($n = 48$, 93.8% agreement at $\pm 5\%$) for the schema-completeness reduces-extraction-error claim, alongside the H4 conditional finding. The §6 retrofit checklist gains a strengthened recommendation: per-retriever measurement is mandatory, and any benchmark of “schema works” must report both the absolute citation rate (gated by index coverage) and the conditional within-cited rate (where the schema effect manifests). The §8.3 longitudinal proposal gains two concrete predictions to track: as retrievers update, the BuiltIn-driven H1 gap may narrow (per-page index coverage of Orbyt will compete with BuiltIn over time), and the H4 advantage may sharpen (schema-rich routes accrue more retrieval signal than prose blog routes). Both trajectories are now measurable on the same 100-query targeted corpus by re-running every six months.