GTM Bench V1: Benchmarking AI Agents on Structured B2B Research

Specialist Retrieval versus Web Research

Working paper

ZoomInfo GTM AI Team*

Draft, June 30, 2026

Abstract

We measure how much a structured data tool helps an AI agent on common B2B research tasks, such as compiling a list of companies or contacts into a fixed table. A single agent harness (Claude Code, running Opus 4.8, Sonnet 4.6, and Haiku 4.5) is evaluated across five retrieval settings: web search only, the ZoomInfo `gtm` CLI, the ZoomInfo MCP, the Apollo.io CLI, and Exa.ai neural search. Each setting is restricted to its one retrieval tool. We run ten tasks with three trials per model (450 agent runs; 448 completed) and record data coverage (how completely the requested table is filled), completion time, token use, and cost. On company lookups every setting fills the table to a similar degree, but coverage there reflects firmographic completeness rather than whether the returned companies meet each task's selection constraint, so we do not read it as parity. The settings diverge where coverage is informative, on contact data, and most of all on verified work email: the specialist data tools fill 84–94% of email cells, while web and neural search fill 22–27%. The data tools also finish faster and consume far fewer tokens. GTM Bench v1 reports coverage rather than correctness. We document one consequence of that choice, a case where a coverage metric ranks a less accurate configuration higher, and we release the full task suite with this paper.

1 Introduction

Sales and marketing teams now point AI agents at research questions that used to be manual analyst work. "Find 50 NYC companies that just raised funding, with headcount and whether they are hiring sales." "Find 20 AI CTOs in the Bay Area with email and LinkedIn." Two things decide whether an agent answers these well: the model's reasoning, and the data tools it can reach. This paper isolates the second factor. Given the same agent and the same prompt, how much does the retrieval tool matter?

We compare web research against four retrieval tools. Two are ZoomInfo's (the `gtm` CLI and the ZoomInfo MCP). We also include Apollo.io as a second structured-data provider and Exa.ai as a neural search system, so the comparison is specialist retrieval against web research in general, not a single-vendor matchup. Where one structured provider is stronger or weaker than another, we report it, but the question of interest is what structured retrieval adds over the open web.

*This study was conducted by ZoomInfo. Two of the evaluated tools are ZoomInfo products (the `gtm` CLI and the ZoomInfo MCP). The task suite covers B2B prospecting, a domain where specialist data tools are expected to perform well, so we treat it as a product-relevant evaluation. To support outside scrutiny we release every task definition (Appendix A) and describe the methodology in full.

This is a working paper. GTM Bench v1 reports *coverage*: how completely each setting fills the requested table, alongside time, token, and cost telemetry. It does not yet verify that returned values are correct. Section 6 explains why this matters and why we still consider the headline result reliable. The task definitions are released with this paper (Appendix A).

2 Methodology

2.1 Arms

An *arm* is an agent configuration that differs from the others only in which retrieval tool it may use. Every arm shares the same model, the same prompt, the same output contract (Section 2.3), file I/O, and a common set of safe local shell utilities such as `jq` and `grep` for processing results. No arm may spawn sub-agents, load project-specific skills, or reach the network except through its assigned tool. Table 1 lists the five arms.

Arm	Retrieval tool	Role
`web`	WebSearch / WebFetch	Baseline: open-web research only
`cli`	ZoomInfo `gtm` CLI	ZoomInfo via its official command-line tool
`zoominfo-mcp`	ZoomInfo MCP	Same backend as `cli`, via native MCP tool-calls
`apollo`	Apollo.io CLI	A second structured-data provider
`exa`	Exa.ai MCP	A neural web-search system

Table 1: The five arms. The only variable across arms is the retrieval tool.

2.2 Isolation guarantees

Because the study rests on each arm using only its assigned tool, we enforce isolation in the harness rather than by instruction:

- **Single retrieval source.** Each arm denies every other arm's tool. The `web` arm cannot invoke `gtm`, `apollo`, or any MCP; the data arms cannot invoke WebSearch or WebFetch. All arms also deny network shell (`curl`, `wget`) and script interpreters.
- **No skill or project-context leakage.** Each run executes in an isolated working directory outside the source repository, with project skills disabled, so no arm inherits vendor-specific guidance that the others lack.
- **No sub-agents.** The sub-agent tool is denied for every arm. Each arm runs as a single agent, which keeps token accounting comparable across arms.
- **Isolated scratch space.** Each run has its own temporary directory, so concurrent runs cannot collide on shared paths.

2.3 Output contract

Every task prompt carries the same contract: write results to `result.csv` with an exact, pre-specified header; one record per row; leave a cell empty when the value cannot be found, and do not guess; aim for the target record count. The instruction to leave cells empty rather than guess matters for how we read the results (Section 6).

2.4 Scoring

GTM Bench v1 scores by fill-rate (coverage):

$$\text{coverage} = \min\left(\frac{\text{filled required cells}}{\text{target\_count} \times \#\text{required columns}},\ 1\right).$$

This penalizes both returning too few rows and leaving required cells empty, measured against the target rather than against whatever an arm chose to return. Optional columns are reported separately and excluded from coverage. Coverage measures whether a cell was filled, not whether the filled value is correct. Section 6 treats that distinction directly.

2.5 Models, trials, and telemetry

Each task and arm is run under three models (Claude Opus 4.8, Sonnet 4.6, Haiku 4.5) and three trials, giving a $10 \times 5 \times 3 \times 3$ matrix of 450 single-shot headless agent runs; 448 completed (two `web`/Sonnet trials of one task produced no output). Per-cell metrics are aggregated by median across trials. Cost and token counts are taken from Claude's per-run result metadata, that is, the input, output, and cache tokens and the dollar cost the model reports for each task.

3 Tasks

The ten tasks (Table 2; full prompts in Appendix A) cover company discovery, contact discovery, enrichment, and buyer-intent. Each task is bounded by a finite, checkable target and defined by a fixed output schema, so coverage is objective. The suite includes routine lookups where we expect little separation between arms, together with contact tasks where a verified work email is the field that distinguishes the arms.

Task	Category	Target
NYC companies funded in last 90 days	company-discovery	50
Boston SaaS companies running Salesforce	company-discovery	30
Fast-growing US fintech by headcount	company-discovery	40
Enrich 20 companies with firmographics	enrichment	20
Companies showing buyer intent on cloud migration	intent	25
IT decision-makers at named accounts (ABM)	contact-discovery	15
AI CTOs in the Bay Area (email + LinkedIn)	contact-discovery	20
VPs of Marketing at Series-B SaaS	contact-discovery	20
Recently-promoted sales leaders	contact-discovery	20
Enrich 15 named contacts	enrichment	15

Table 2: The ten tasks. Five are company-side and five are person-side.

4 Results

4.1 What coverage does and does not show on company tasks

Table 3 gives mean coverage per arm per model, and Table 4 breaks it out by category. On the company-discovery tasks, and on the intent task, coverage is high for every arm, including `web`. This result needs care, because on these tasks coverage measures less than it appears to.

The scored columns on the company tasks are firmographic: company name, website, employee count. Every real company has these, and a web researcher can find them. The property that actually defines each task, namely funded in the last 90 days, currently running Salesforce, growing quickly, or in-market for cloud migration, is the selection constraint, and it is not a scored column. The `nyc-funding-90d` task, for example, scores only name, website, and headcount, and contains no field that records when a company raised money. An arm can return a full table of plausible companies and reach high coverage without those companies satisfying the constraint, and coverage does not check that they do.

The `saas-using-salesforce-boston` task shows the gap concretely. Both `web` and `cli` return 30 companies and score 100%, but on different grounds, and they return largely different companies. The `web` arm infers Salesforce use from indirect signals such as a company advertising a "Salesforce Administrator" role; the `cli` arm cites ZoomInfo technographic tech-install records. Fill-rate treats the two as equal. Whether the returned companies actually use Salesforce and sit in Boston is what a structured tech-install filter establishes, and is what coverage cannot see. We therefore do not read the company-discovery numbers as evidence that web research is sufficient for these tasks. That question turns on constraint satisfaction, which the correctness work in Section 7 will measure. The separation we can report from coverage alone is on contact data.

Model	`cli`	`zoominfo-mcp`	`apollo`	`exa`	`web`
Opus 4.8	97	99	100	87	85
Sonnet 4.6	98	98	98	93	92
Haiku 4.5	99	99	82	89	88

Table 3: Mean coverage (%) across all ten tasks, by arm and model.

Category (Opus)	`cli`	`zoominfo-mcp`	`apollo`	`exa`	`web`
Company-discovery (4 tasks)	100	100	100	100	98
Buyer-intent (1 task)	100	100	100	100	100
Contact (5 tasks)	94	97	99	74	71

Table 4: Mean coverage (%) by task category, Opus. The arms separate on contact tasks.

4.2 The headline: contact data

The separation comes from contact tasks, and within those from the verified work-email field. Figure 1 shows email-column coverage across the five contact tasks. The specialist data tools fill 84–94% of email cells. Web and neural search fill 22–27%, because open-web research generally cannot return a verified work email, and much of the 22–27% it does return is likely a pattern guess of the form `first.last@domain`. This gap reflects what each tool can supply, so it does not depend on the coverage-versus-correctness question discussed in Section 6.

On contact discovery (for example, AI CTOs in the Bay Area), the data arms reach 98%–100% coverage while `web` and `exa` sit at 67%–83%, and the shortfall is concentrated in the email column.

4.3 Cost, tokens, and time

The data tools are also more efficient. Table 5 aggregates Opus cost, work tokens (input plus output plus cache-creation; cheap cache reads excluded), and total agent time across the ten tasks. The `web`

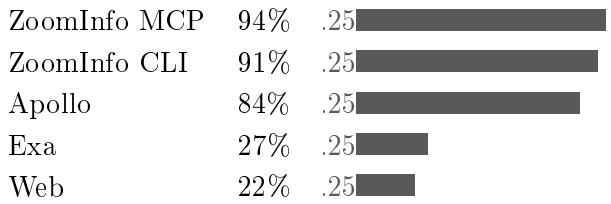

Figure 1: Verified work-email coverage across the five contact tasks (median over models and trials). Specialist data tools supply emails at roughly four times the rate of web or neural search.

baseline costs about four times the ZoomInfo CLI and consumes more than twenty times the work tokens, because it issues many searches and fetches and reads large pages, whereas a structured query returns compact records. The ZoomInfo MCP records the lowest total time.

Arm (Opus, 10 tasks)	Cost (USD)	Work tokens	Total time
`cli`	$10.12	0.46M	53 min
`apollo`	$14.78	0.67M	101 min
`zoominfo-mcp`	$15.80	1.48M	32 min
`exa`	$36.75	3.38M	55 min
`web`	$43.88	11.10M	102 min

Table 5: Cost, work tokens, and total agent time across the ten tasks (Opus), as reported by Claude's per-run metadata.

4.4 Specialist retrieval versus the field

Taking coverage, cost, and stability together: on Opus the ZoomInfo arms match or beat the best other arm on coverage for 6 of 8 comparable tasks, and at lower cost. Apollo is even on coverage on Opus and Sonnet, but it is the costlier contact path, its per-person email reveal is rate-limited, and its mean coverage falls to 82 on Haiku (and to zero emails on the contact-enrichment task) when the smaller model cannot work around that limit. The ZoomInfo CLI holds high coverage, the lowest cost, and stable behavior across all three models.

5 Analysis

5.1 Case study: acquired-company resolution

The clearest within-category split among the structured tools appears on `enrich-contact-list`, which gives 15 people by role and company. Apollo reaches 97% coverage on Opus while the ZoomInfo CLI reaches 73%. We looked into the gap. It is mainly a query-resolution problem, not a ZoomInfo data gap:

- Two of the people work at acquired companies (Segment is now part of Twilio, Drift is now part of Salesloft). Apollo's people-search maps the brand to its current parent on its own. A query to the `gtm` CLI by the brand name returns unrelated companies, and the agent did not perform the acquisition mapping, so it left the cells empty as the contract requires.
- Direct probes confirm that ZoomInfo holds these people with verified emails once the query uses the parent company. The shortfall is reachability and agent strategy, not missing data.
- The remaining misses are a mix of genuine coverage gaps and one stale affiliation.

We report this result in full because an honest comparison should show where each tool is weaker as well as stronger.

5.2 A coverage metric can rank a less accurate setting higher

On `enrich-contact-list` the ZoomInfo CLI's coverage rises as the model gets smaller: Opus 73%, Sonnet 85%, Haiku 95%. Inspection shows this is an artifact of fill-rate scoring, not a quality gain. Opus recognizes the acquired-company ambiguity and leaves the cells empty. Haiku takes the top fuzzy match and enriches the wrong company. For "VP of Engineering, Segment" it returns an email at an unrelated firm also named Segment; for "CRO, Drift" it returns a contact at a different Drift. Because coverage does not check correctness, Haiku's filled-but-wrong cells score higher than Opus's empty-but-correct ones. The more careful model scores lower.

This is a general caution for agent benchmarks. A completeness metric on its own can rank a less accurate configuration above a more accurate one. We report it as a result, and it sets up the correctness work described in Section 7.

6 Limitations

1. **Coverage is not correctness.** GTM Bench v1 measures whether the requested cells were filled, not whether the values are right. Section 5.2 shows that this can change a ranking. We therefore limit strong claims to where coverage and capability coincide: the contact-email gap against web and neural search (Figure 1), where the low baseline number reflects an inability to return verified emails rather than caution. The company-discovery and intent tasks are the most affected by this limitation: their selection constraint (recent funding, a tech install, growth, in-market intent) is not a scored column, so high coverage there does not imply the returned companies meet the constraint (Section 4). Comparisons among the structured providers, and among models, are reported as coverage only and should be read as provisional pending the correctness work in Section 7.
2. **Authorship and task selection.** ZoomInfo authored the benchmark, and the tasks target B2B prospecting. We release every task definition (Appendix A) so readers can judge task selection for themselves.
3. **Single run.** Live data and non-deterministic agents make the results a point-in-time snapshot. 448 of 450 cells completed.
4. **Harness.** The harness and scoring code are not released with this version. The methodology is described in enough detail to reproduce in principle.

7 Future work

GTM Bench v2 will pair the agent coverage measured here with correctness. ZoomInfo already operates large-scale data-quality systems and a data-research team of roughly 300 analysts who continuously verify the accuracy of its data. v2 will connect the harness's coverage testing to these existing verification practices, so that completeness and correctness are reported together. As Section 5.2 shows, correctness scoring should reduce artifacts such as the Haiku result and give a truer picture of each setting. v2 will also broaden the task suite and add further retrieval tools.

8 Conclusion

Across ten structured B2B research tasks and three models, coverage on company lookups is similar across settings, but that similarity reflects what fill-rate measures, namely firmographic completeness, rather than whether the returned companies satisfy each task's selection constraint. The constraint is where structured filters are grounded, and v1 does not score it. The separation that coverage can show cleanly is on contact data: specialist tools fill 84–94% of verified work emails against 22–27% for web and neural search, at roughly a quarter of the cost and a fraction of the tokens. Among the specialist tools, the ZoomInfo CLI combines high coverage, the lowest cost, and stable behavior across model sizes. These are v1 coverage results, and correctness verification, which we expect to widen the gap on the company tasks, is the next step.

A Task definitions

The verbatim prompt and required and optional columns for each task. Every prompt runs with the common output contract of Section 2.3.

NYC companies that raised funding in the last 90 days

id: `nyc-funding-90d` *category:* company-discovery *target:* 50

```
Find 50 companies headquartered in the New York City metro area that raised
a funding round in the last 90 days (as of the run date). For each company I
need its website, current employee count, and whether it is currently hiring
for sales roles.
```

Required columns: company_name, website, employee_count. **Optional:** hiring_sales_now.

Boston SaaS companies running Salesforce

id: `saas-using-salesforce-boston` *category:* company-discovery *target:* 30

```
Find 30 software/SaaS companies headquartered in the Boston metro area that
currently use Salesforce as their CRM. For each, give the company name, its
website, its approximate employee count, and the evidence that they use
Salesforce (e.g. tech-install record, job posting, case study).
```

Required columns: company_name, website, employee_count. **Optional:** salesforce_evidence.

Fast-growing US fintech companies by headcount

id: `high-growth-fintech-headcount` *category:* company-discovery *target:* 40

```
Find 40 fintech companies in the United States with between 50 and 1,000
employees that are growing quickly (meaningful recent headcount growth). For
each, give the company name, its website, its current employee count, and a
growth signal (e.g. recent headcount change, recent funding, or hiring surge).
```

Required columns: company_name, website, employee_count. **Optional:** growth_signal.

Enrich a list of 20 companies with firmographics

id: `enrich-company-list` *category:* enrichment *target:* 20

```
Enrich the following 20 companies. For each, return its official website,
industry, headquarters city + state/country, and current employee count.

Companies: Stripe, Databricks, Snowflake, Plaid, Notion, Figma, Ramp,
Brex, Airtable, Retool, Vercel, Linear, Rippling, Gusto, Carta, Mercury,
Scale AI, Anduril, Wiz, Canva.
```

Required columns: company_name, website, industry, headquarters, employee_count. **Optional:** (none).

Companies showing buyer intent on cloud migration

id: `intent-cloud-migration` *category:* intent *target:* 25

```
Find 25 US companies with 250+ employees that are currently showing active
buyer intent / in-market signals around cloud migration or cloud
infrastructure. For each, give the company name, its website, its employee
count, and the intent signal or evidence (topic and recency if available).
```

Required columns: company_name, website, employee_count. **Optional:** intent_signal.

IT decision-makers at named target accounts (ABM)

id: `abm-account-decision-makers` *category:* contact-discovery *target:* 15

```
For each of these 5 companies - Cloudflare, Datadog, MongoDB, Okta, and
HashiCorp - find 3 senior IT / engineering decision-makers (Director level or
above, e.g. CIO, CTO, VP Engineering, VP IT). For each person give their full
name, title, company, work email address, and LinkedIn profile URL.
```

Required columns: full_name, title, company, email. **Optional:** linkedin_url.

AI CTOs in the Bay Area with email + LinkedIn

id: `bay-ai-ctos` *category:* contact-discovery *target:* 20

```
Find 20 CTOs at AI companies in the San Francisco Bay Area. For each person I
need their full name, the company they work at, their work email address, and
their LinkedIn profile URL.
```

Required columns: full_name, company, email. **Optional:** linkedin_url.

VPs of Marketing at Series-B SaaS companies

id: `vp-marketing-seriesb-saas` *category:* contact-discovery *target:* 20

```
Find 20 people with a VP of Marketing (or equivalent VP-level marketing) title
at US-based Series-B-stage SaaS companies. For each, give their full name, job
title, company, work email address, and LinkedIn profile URL.
```

Required columns: full_name, title, company, email. **Optional:** linkedin_url.

Recently-promoted sales leaders (job-change signal)

id: `recently-promoted-sales-leaders` *category:* contact-discovery *target:* 20

```
Find 20 people in the US who were recently promoted into or recently started a
senior sales leadership role (e.g. VP Sales, CRO, Head of Sales) within roughly
the last 90 days. For each, give their full name, current title, company, work
email address, LinkedIn profile URL, and the change signal (what changed and
roughly when).
```

Required columns: full_name, title, company, email. **Optional:** linkedin_url, change_signal.

Enrich a list of 15 named contacts

id: `enrich-contact-list` *category:* enrichment *target:* 15

```
Enrich the following 15 people. For each, return their current job title,
current company, work email address, and LinkedIn profile URL. Use the
name + company hint provided to disambiguate.

1. Head of Demand Generation, HubSpot
2. VP of Product Marketing, Zendesk
3. Director of Revenue Operations, Gong
4. VP of Sales, Outreach
5. Chief Marketing Officer, Klaviyo
6. Head of Growth, Webflow
7. VP of Engineering, Segment
8. Director of Partnerships, Asana
9. CRO, Drift
10. VP of Customer Success, Intercom
11. Head of People, Lattice
12. VP of Finance, Calendly
13. Director of Product, Miro
14. Chief Information Officer, DocuSign
15. VP of Marketing, Amplitude
```

Required columns: name_or_role, title, company, email. **Optional:** linkedin_url.

B Full coverage scoreboard

Coverage (%) per task and arm, by model. Median over trials.

Opus 4.8

Task	cli	zi-mcp	apollo	exa	web
nyc-funding-90d	100	100	100	100	91
saas-using-salesforce-boston	100	100	100	100	100
high-growth-fintech-headcount	100	100	100	100	100
enrich-company-list	100	100	99	100	100
intent-cloud-migration	100	100	100	100	100
abm-account-decision-makers	100	100	100	78	75
bay-ai-ctos	98	100	100	68	67
vp-marketing-seriesb-saas	100	100	100	75	75
recently-promoted-sales-leaders	100	100	100	75	75
enrich-contact-list	73	85	97	75	63

Sonnet 4.6

Task	cli	zi-mcp	apollo	exa	web
nyc-funding-90d	100	100	99	100	97
saas-using-salesforce-boston	100	100	100	100	100
high-growth-fintech-headcount	100	100	100	100	100
enrich-company-list	100	100	96	100	100
intent-cloud-migration	100	100	100	100	100
abm-account-decision-makers	100	100	100	98	100
bay-ai-ctos	100	100	100	82	83
vp-marketing-seriesb-saas	100	100	100	99	100
recently-promoted-sales-leaders	99	96	100	75	75
enrich-contact-list	85	82	85	78	67

Haiku 4.5

Task	cli	zi-mcp	apollo	exa	web
nyc-funding-90d	100	100	66	82	95
saas-using-salesforce-boston	100	100	67	100	100
high-growth-fintech-headcount	100	100	78	100	100
enrich-company-list	100	100	99	100	100
intent-cloud-migration	100	100	67	100	100
abm-account-decision-makers	98	100	90	80	77
bay-ai-ctos	97	100	100	100	70
vp-marketing-seriesb-saas	100	100	100	75	83
recently-promoted-sales-leaders	95	100	75	80	78
enrich-contact-list	95	90	75	75	73

References

[1] Anthropic. *Claude Code*. `https://claude.com/claude-code`, 2026.

[2] ZoomInfo GTM AI Team. *GTM Bench task suite (v1)*. Released with this paper, 2026.