# Introducing $C^4$ and Ebla

*Benchmarking and Reinforcement Learning for Grounded Enterprise Reasoning*

**Abstract.** Knowledge workers routinely search for, synthesize, and verify information scattered across heterogeneous internal documents—PDFs, spreadsheets, and slide decks. Despite rapid progress in frontier language models, AI agents remain far from matching human performance on these tasks. We introduce $C^4$, a benchmark for grounded reasoning—retrieval, synthesis, and verification over heterogeneous internal documents—in realistic enterprise settings. $C^4$ comprises 40 tasks distributed across 3 independent simulated enterprise environments, each generated from a comprehensive knowledge graph. Tasks are authored by domain specialists and evaluated along four independently scored axes—**Correctness**, **Completeness**, **Composition**, and **Citations**—with signed weights that penalize fabrication. We evaluate seven frontier models, all at high reasoning effort, across 840 total runs ($n$=3 per task–model pair). The best frontier model achieves 20.1%; only 6.1% of task–model pairs are fully solved—underscoring the benchmark's difficulty. After reinforcement fine-tuning on 30 tasks using the $C^4$ rubric score as reward, a 120B open-weight model (Ebla) achieves 25.4% on the full 40-task benchmark—surpassing all frontier models. The multi-axis rubric provides dense, decomposed reward signal suitable for reinforcement learning, and the closed, contamination-free environments ensure reproducibility.

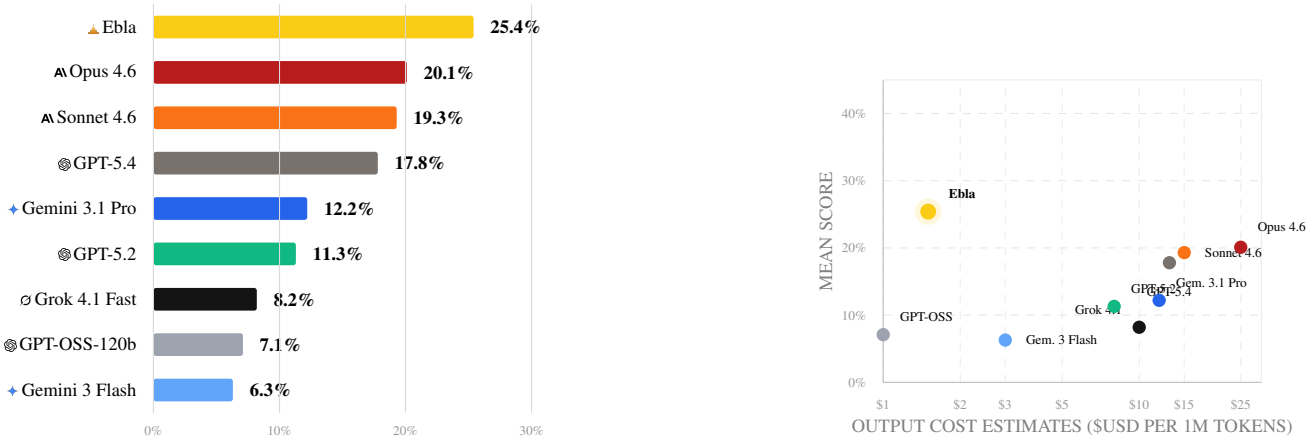



**Figure 1.** (a) Overall $C^4$ scores (%). All models at high reasoning effort. (b) Performance vs. output cost estimates ($USD per 1M tokens).

## 1 Introduction

Enterprise knowledge work requires navigating a complex information landscape: policy documents contradict training materials, financial figures span multiple spreadsheets, and critical details are embedded in slide decks rather than searchable text. A single business question may require retrieving documents across departments, parsing visual elements such as charts and tables, chaining inferences across sources, and calibrating confidence when evidence is incomplete or absent.

Recent benchmarks have begun to address document-grounded question answering [Salesforce Research, 2025, Databricks, 2025] and conversational retrieval [IBM Research, 2025], but they typically evaluate on curated, text-only corpora with single-score metrics. This leaves several dimensions of real enterprise search untested: native visual understanding of PDF layouts, reasoning over complex work-in-progress documents, and the ability to abstain or qualify answers when evidence is insufficient.

We introduce $C^4$, a benchmark designed to close this gap. $C^4$ comprises three independent simulated organizations: Zyro, a SaaS analytics platform; Vexel, a financial services firm; and Korvo, a chemical manufacturing company. Each is generated from a comprehensive knowledge graph that populates both a document corpus and high-fidelity platform clones (e.g., Salesforce, ServiceNow, Workday). Documents include work-in-progress drafts, training materials, and org charts—the heterogeneity of real enterprise data. Tasks are evaluated along four independently scored axes—Correctness, Completeness, Composition, Citations—with signed weights per criterion. Fabrication penalties can drive individual axis scores below zero, and the resulting decomposition surfaces distinct failure modes that a single aggregate score would obscure.

The rubric's fine granularity also makes $C^4$ a productive training environment: the four independently scored axes provide decomposed reward dimensions amenable to curriculum-based reinforcement learning, and the signed-weight criteria produce near-continuous reward distributions rather than sparse binary signal. We demonstrate this by RL fine-tuning GPT-OSS-120b (a 120B open-weight

model) on 30 tasks using OpenAI's RFT API with the $C^4$ rubric score as reward, producing **Ebla**. On the full 40-task benchmark, Ebla scores 25.4%—5 pp above the best frontier model.

Across frontier evaluations, we identify failure modes—failure to abstain when the corpus lacks the requested information, visual misinterpretation of charts and diagrams, and cross-document arithmetic fabrication—that represent structural limitations beyond what scaling alone is likely to resolve.

We evaluate seven frontier models—Opus 4.6, Sonnet 4.6, GPT-5.4, GPT-5.2, Gemini 3.1 Pro, Grok 4.1 Fast, and Gemini 3 Flash—all at high reasoning effort, using a deep research agent (Figure 2) that follows a search-fetch-answer loop over each environment's corpus. Figure 1 summarizes overall results: the best frontier model reaches 20.1%. Only 6.1% of task–model pairs are fully solved. After RL training, Ebla reaches 25.4% at the lowest inference cost.

## 2 Related Work

Several recent benchmarks address enterprise and document-grounded reasoning. Table 1 positions $C^4$ relative to the most relevant prior work.

**HERB** [Salesforce Research, 2025] evaluates enterprise artifact search across Slack messages, documents, and pull requests using text-only simulated logs without visual document understanding. **WixQA** [Wix Research, 2025] targets support QA over a polished help-center knowledge base, offering limited multi-document reasoning. **OfficeQA** [Databricks, 2025] provides grounded reasoning on U.S. Treasury PDFs with agentic toolchains, but operates over government bulletins—clean, structured documents that differ markedly from enterprise reality. **MTRAG** [IBM Research, 2025] tests multi-turn conversational retrieval-augmented generation with curated domain texts but collapses evaluation to a single score. **UAEval4RAG** [Salesforce Research, 2024] focuses narrowly on unanswerable query synthesis and detection. **KARLBench** [Databricks AI Research, 2026] aggregates six existing search benchmarks—BrowseComp-Plus, TREC-Biogen, FinanceBench, QAMPARI, FreshStack, and an internal PMBench—into a multi-capability evaluation suite and demonstrates strong RL-trained agents via multi-task off-policy reinforcement learning. KARLBench operates over text-only vector search without visual document understanding and uses nugget-based single-score evaluation rather than decomposed multi-axis rubrics.

$C^4$ differs from all of the above in two structural ways. First, every corpus is authored by domain experts rather than assembled from existing benchmarks or scraped from public sources; this eliminates contamination risk by construction rather than by deduplication. Second, $C^4$ requires the conjunction of capabilities that no single prior benchmark tests: native visual PDF understanding of charts, tables, and images; reasoning over complex, heterogeneous documents including work-in-progress drafts; three independent closed corpora cross-modal multi-hop reasoning; and decomposed multi-axis, trajectory-level evaluation with fabrication penalties.

## 3 Methodology

### 3.1 Environment Construction

$C^4$ comprises three independent simulated enterprise environments—Zyro, a SaaS analytics platform; Vexel, a financial services firm; and Korvo, a chemical manufacturing company—each representing a distinct organization, industry vertical, and document style. The construction pipeline (Figure 3) proceeds in four stages: (1) knowledge graph construction, (2) document corpus generation, (3) task authoring, and (4) gold output production. Each stage includes iterative review loops with domain-expert sign-off before advancing.

**Knowledge Graphs.** A comprehensive knowledge graph encodes entities (people, teams, documents, products, customers, certifications, financial records) and their relationships for each environment, serving as the single source of truth from which all downstream artifacts are derived. Figure 4 shows the graph composition for each environment. The benchmark includes 40 tasks (8 Zyro, 11 Vexel, 21 Korvo), many requiring cross-domain reasoning—for example, calculating a team's budget utilization by joining an org chart, a compliance report, and a financial spreadsheet.

**Expert Network.** Environment construction is performed by a network of domain experts (Figure 5)—senior accountants, management consultants, compliance officers, software engineers, and operations specialists, each with 5–15 years of industry experience—organized across six primary industry clusters: technology and enterprise platforms, financial services, consulting and advisory, government and regulators, industrial and aerospace, and people/HR/GRC systems. Six specialized roles collaborate across the pipeline (shown in Figure 3): document authors create the corpus, document reviewers validate narrative consistency, task authors design evaluation tasks, task reviewers QA rubrics for solvability, platform experts specify clone interfaces, and clone engineers build the high-fidelity application mockups.

**Document Corpus.** Both authored documents and high-fidelity platform clones—internal replicas of enterprise tools such as Salesforce, ServiceNow, and Workday—are populated directly from the knowledge graph. The corpus spans **DOCX**, **PPTX**, and **XLSX** artifacts (including `.xlsx`, `.xlsm`, and `.csv`). Zyro and Vexel resemble broad departmental file systems (e.g., analytics, compliance, training; finance-billing, strategy-planning, technology-devops), whereas Korvo is a tighter contract repository of exhibits, schedules, appendices, and the MSA. Modality differs as well: Zyro is slide-heavy, Vexel mixes slides with spreadsheets, and Korvo contains no PPTX and is markedly more text-dense. Figure 6 shows the format breakdown for each environment.

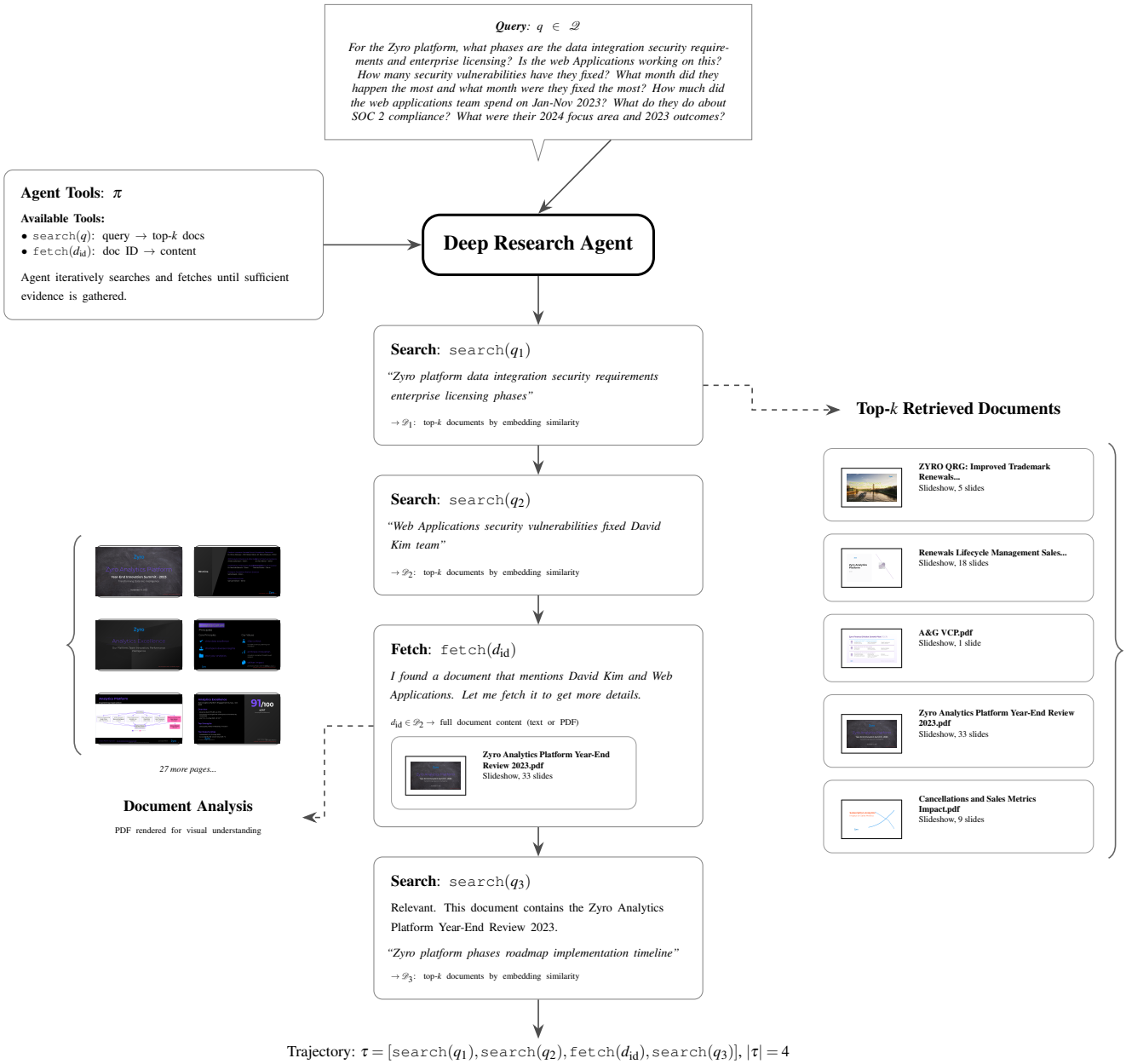



**Figure 2.** Deep Research Agent pipeline. Given a natural-language query $q$, the agent iteratively issues `search` calls against a document corpus indexed by `text-embedding-3-large` + pgvector, retrieves the top-$k$ results, and selectively `fetch`es individual documents for analysis. For PDF documents, an isolated LLM call to the native PDF API produces a structured analysis without injecting raw document tokens into the main context. The agent terminates when it judges that sufficient evidence has been gathered, and submits a final response with page-level citations.

**Table 1.** Comparison of $C^4$ with related enterprise and document benchmarks. ✓ = full support, ∼ = partial, × = absent.

| | Visual | Complex docs | Closed corpus | Agentic | Multi-hop | Multi-axis rubric |
|---|---|---|---|---|---|---|
| HERB [Salesforce Research, 2025] | × | × | ✓ | ✓ | ✓ | × |
| WixQA [Wix Research, 2025] | × | × | ✓ | × | ∼ | × |
| OfficeQA [Databricks, 2025] | ∼ | × | ✓ | ✓ | ✓ | ∼ |
| MTRAG [IBM Research, 2025] | × | × | ✓ | ∼ | ✓ | × |
| UAEval4RAG [Salesforce Research, 2024] | × | × | ✓ | × | × | × |
| KARLBench [Databricks AI Research, 2026] | × | ∼ | ✓ | ✓ | ✓ | × |
| **$C^4$ (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |


**Figure 3.** Environment construction pipeline. Each environment is built through four phases with iterative review loops. The knowledge graph serves as the single source of truth from which all downstream artifacts—documents, platform clones, tasks, and gold outputs—are derived.

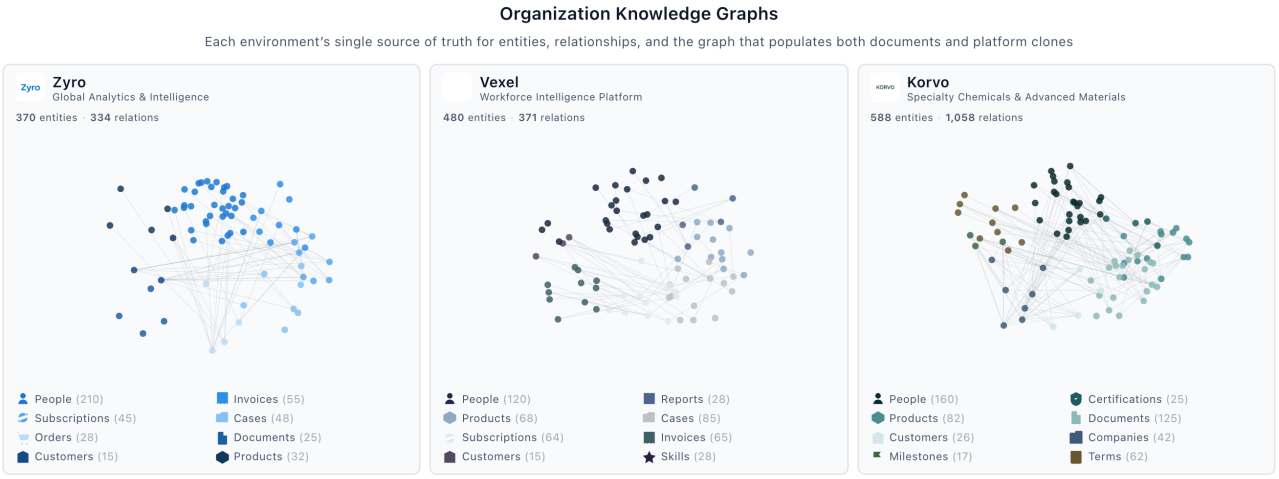



**Figure 4.** Knowledge graph composition for the three $C^4$ environments. Circle size is proportional to entity count; edges represent typed relationships. Korvo's high relation-to-entity ratio (1.8×) reflects a densely interconnected supply-chain domain.

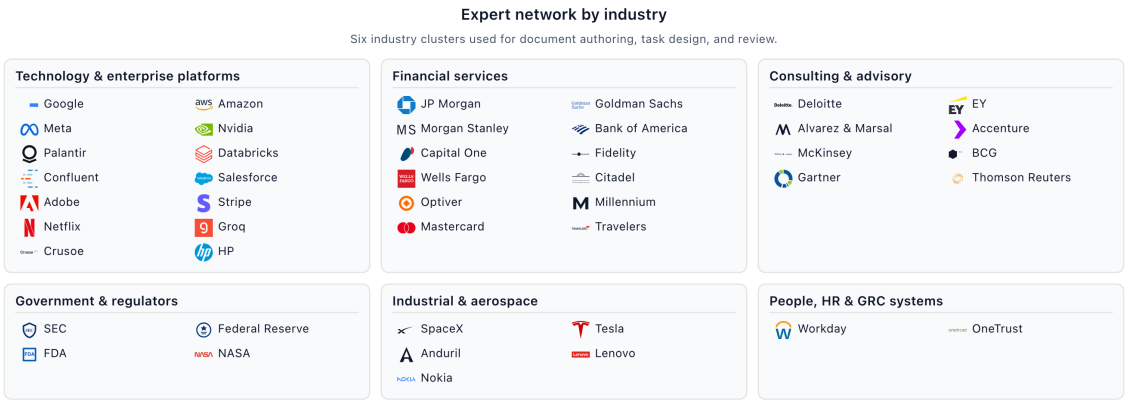



**Figure 5.** Select experts from the domain specialist network. The full network is organized into six industry clusters to ensure occupational realism across technology, finance, advisory, public-sector, industrial, and workflow-heavy enterprise environments.

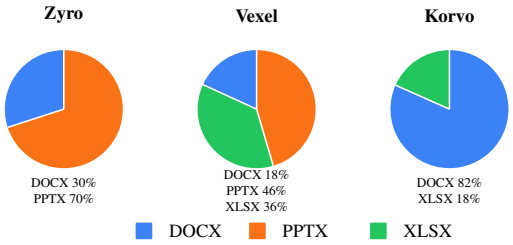



**Figure 6.** Corpus composition by environment: Zyro 6 DOCX / 14 PPTX; Vexel 4 DOCX / 10 PPTX / 8 XLSX; Korvo 49 DOCX / 11 XLSX.

**Task Design.** Domain experts write long-horizon tasks that can only be solved by navigating the environment's files and tools. For each task, experts specify a single-turn natural-language prompt, the required output format, and a rubric of binary evaluation criteria. Tasks are calibrated across four difficulty levels (Figure 7), from single-document lookups (<30 min human time) to multi-step expert reasoning requiring 6+ documents and deep domain expertise (3+ hours). Experts then perform each task themselves to produce gold outputs, verifying each against the rubric to confirm alignment.

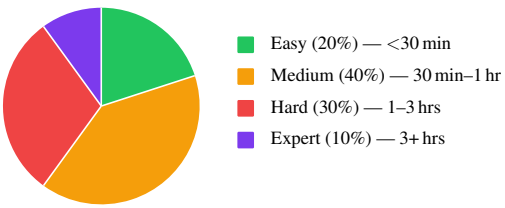



**Figure 7.** Task difficulty distribution. The benchmark is weighted toward medium and hard tasks to maximize discriminative power, with easy and expert tasks anchoring the endpoints.

### 3.2 Agent

Following the pattern of deep research agents [OpenAI, 2024a], we equip the agent with two tools:

- `search`$(q)$: Queries the full document corpus and returns up to $k$=5 results ranked by cosine similarity over `text-embedding-3-large` embeddings (pgvector). Each result includes the document ID, title, and a short excerpt (∼250 characters). Excerpts are sufficient to judge relevance but not to answer the query, forcing the agent to fetch for full content.
- `fetch`$(d_{\text{id}}, q)$: Retrieves and analyzes a single document. For PDFs, an isolated LLM call to the model's native PDF API analyzes the document against the query and returns a structured text analysis; the raw PDF never enters the main context window. This design enables visual understanding of charts, tables, and images without the token cost of serializing multi-page documents.

The agent operates in a search-fetch-answer loop under a fixed budget of 36 tool-use steps: it formulates search queries, examines results, fetches promising documents, and continues until it judges that sufficient evidence has been gathered or the budget is exhausted. It then submits a final response with page-level citations specifying the document ID and page number(s) for each claim. Figure 2 illustrates the full pipeline with an example trajectory.

### 3.3 Infrastructure

$C^4$ was developed in collaboration with HUD, an AI evaluation platform that provides the infrastructure for executing and scoring agent trajectories at scale. HUD contributed to environment deployment, scenario orchestration, and quality assurance of the evaluation pipeline, ensuring that scoring was reproducible across thousands of runs. Each environment runs in an isolated Docker container with all documents synced from external storage at build time and preloaded at startup, ensuring identical initial state across runs with no runtime dependencies. The document corpus is indexed at build time using `text-embedding-3-large` embeddings stored in PostgreSQL with pgvector, so all retrieval operates over a fixed, pre-computed index. This infrastructure supports both batch evaluation and online RL training.

### 3.4 Evaluation

$C^4$ evaluates agent outputs along four independently scored axes: **Correctness** (*Cr*), **Completeness** (*Cm*), **Composition** (*Co*), and **Citations** (*Ci*). Each task has a rubric of $N$ binary criteria partitioned across axes $\mathscr{A} = \{Cr, Cm, Co, Ci\}$. Let $v_i \in \{0,1\}$ denote the LLM judge's verdict for criterion $i$ and $w_i$ its signed weight (positive for rewards, negative for penalties such as fabrication). The final score is:

$$R_{\text{final}} = \frac{\sum_{k \in \mathscr{A}} \sum_{i \in k} v_i \cdot w_i}{\sum_{i : w_i > 0} w_i} \tag{1}$$

The weight budget is allocated roughly in the order Correctness > Completeness > Citations > Composition. Each criterion is evaluated via a separate single-shot LM Judge call using Gemini 3.1 Pro. In our assessment of judge reliability, Gemini 3.1 Pro matched the expected verdict on 99% of criteria; we observed one error case in which it marked a criterion as unmet when it should have been marked met. Because the axes are scored independently, the framework surfaces distinct failure modes that a single aggregate score would conflate.

**Correctness** accounts for the largest share of the rubric weight. Criteria map one-to-one with verifiable facts extracted from the gold output, and carry fabrication penalties: a confidently hallucinated value incurs a negative weight, so a wrong answer can score below zero on this axis.

**Completeness** captures cross-document synthesis and multi-hop reasoning. An agent that retrieves one source but misses the second corroborating document loses Completeness credit even when its Correctness on found items is high.

**Composition** evaluates structure, terminology, formatting, and efficiency. Agents that take fewer steps to reach a correct answer receive higher Composition scores; efficiency is rewarded within this axis rather than applied as a global penalty that would scale down the other axes. This is the most uniform axis across models—all frontier models produce well-structured output—making it a useful diagnostic: high Composition paired with low Correctness signals overconfident answering.

**Citations** measures grounding quality. Models vary most here in their ability to provide specific document IDs and page numbers rather than vague references. Citation accuracy correlates with Correctness but is not redundant: an agent can extract the right value from the right document but cite the wrong page, or cite correctly but misinterpret what it read.

### 3.5 Training

The $C^4$ rubric provides dense partial credit—an RL agent receives gradient signal from correctly retrieved documents, accurately extracted values, and valid citations even when the overall task is unsolved. Each task's rubric contains binary criteria with signed weights, producing a near-continuous reward distribution rather than the sparse binary signal typical of many agentic benchmarks. The four axes provide independent reward dimensions that enable curriculum strategies (e.g., weighting Correctness heavily early, then increasing Completeness and Citations weights as the agent matures). The simulated corpora are contamination-free by construction, and Docker-based execution ensures identical initial conditions across runs; the only per-trace difference is which environment's corpus is mounted for the agent. Because the three environments share no documents or entities, cross-environment transfer during training serves as a natural measure of generalization: improvements on Zyro training tasks that transfer to Korvo evaluation tasks cannot be attributed to memorization.

We trained GPT-OSS-120b, a 120B open-weight model that scores 7.1% on $C^4$—below every frontier model evaluated. Most runs produce near-zero scores: the model either fabricates answers that trigger penalty criteria or hedges so aggressively that it fails to answer the question. By rubric axis, Correctness is 10.1%, Completeness 6.3%, Citations 8.2%, and Composition 16.2% (highest, since even a poor response can be well-formatted).

**Vision backbone selection.** GPT-OSS-120b has no native vision or PDF understanding capability, so the `fetch` tool's document analysis—which for frontier models uses each model's own PDF API—requires a separate vision model. We selected Gemini 3 Flash: despite scoring lowest among frontier models on $C^4$ overall (6.3%), it demonstrates strong document understanding when used as an isolated analysis component. The bottleneck is not perception but orchestration.

A controlled experiment confirmed this. We paired two orchestrator models (Opus 4.6 and Gemini 3 Flash) with the *same* vision backbone (Gemini 3 Flash) for document analysis, holding all other variables constant: identical tool budgets, environments, and evaluation rubrics. Opus 4.6 with Gemini 3 Flash vision consistently outperformed Gemini 3 Flash orchestrating itself on visual tasks—despite identical underlying vision capabilities. Trajectory analysis revealed the mechanism: the stronger orchestrator issued additional `fetch` calls to re-examine pages it had already retrieved and cross-checked extracted values against surrounding context before committing to an answer. The weaker orchestrator accepted its first-pass extraction without revisiting documents, even when extracted values were inconsistent with other retrieved evidence. Performance on multimodal tasks is driven by the orchestrator—the agent deciding what to search, when to re-examine, and how to triangulate—not the vision component it delegates to.

This finding directly shaped the Ebla training strategy: rather than sourcing a stronger vision backbone, we focused RL training entirely on shaping the *orchestration* layer—the search, retrieval, and verification behaviors that surround the vision call. The resulting post-training trajectories con-

firm the approach: Ebla learned to fetch more documents, re-examine pages containing charts or tables, and cross-reference visually extracted data against textual evidence elsewhere in the corpus—the same verification-seeking patterns that distinguished the stronger orchestrator in our controlled comparison.

We used OpenAI's Reinforcement Fine-Tuning (RFT) API [OpenAI, 2024b] with the $C^4$ rubric score as reward. The training set consists of 30 tasks drawn uniformly across all three environments (10 per environment) and stratified by difficulty across anchor tasks (baseline score 60–75%), main tasks (25–60%), and stretch tasks (<25%). Training ran single-epoch, no SFT warmup, no filtering, or rubric modifications. Fabrication penalties were retained in the reward signal.

Each training trace receives the normalized rubric score $R_{\text{final}}$ from Equation 1 as its reward, mapped to $[0,1]$. Because the numerator includes negative-weight criteria, a trace that triggers fabrication penalties can score below a blank response. The per-criterion binary verdicts are produced by a single-shot LM Judge (Gemini 3.1 Pro) operating independently on each criterion, and the agent's full trajectory—including search queries, fetch calls, and the final response—is available to the judge for trajectory-aware criteria such as retrieval efficiency. The agent operates under a fixed budget of 36 tool-use steps per trace, with the same two tools (search and fetch) used during evaluation.

The rubric's negative weights created a predictable training dynamic. Early in training, the model discovered that terse, noncommittal responses avoid fabrication penalties entirely—producing a reward of approximately zero rather than the negative rewards that confident hallucination incurs. Refusal rates climbed from 8% of traces at initialization to over 40% within the first 200 training steps, while mean reward plateaued near zero. Trajectory-level observability surfaced the pattern: the model was generating increasingly hedged responses ("Based on the available documents, I was unable to find sufficient evidence to answer this question") even on tasks where the evidence was clearly retrievable.

We retained our penalty structure for our rubrics. By approximately step 400, the model broke learned that precisely citing what documents contain while explicitly flagging what they *don't* earns substantial positive credit without triggering fabrication penalties. The resulting behavior is qualitatively different from both the base model (which fabricates confidently) and the mid-training refusal mode (which says nothing): Ebla produces calibrated, evidence-grounded responses that earn reward on Correctness and Citations simultaneously. This suggests that penalty-based reward signals, while harder to optimize, can produce more robust epistemic behavior than penalty-free alternatives.

# 4 Results

We evaluate seven frontier models across all 40 tasks with $n$=3 runs per task–model pair (840 total runs). All models are set to high reasoning effort and use the same agent architecture, system prompt, and tool definitions; only the underlying LLM varies.

## 4.1 Benchmark

### 4.1.1 Overall Results

Figure 1 reports aggregate scores. Opus 4.6 leads at 20.1%, followed by Sonnet 4.6 (19.3%) and GPT-5.4 (17.8%). The gap between the best and worst frontier model is 13.8 percentage points. Across 280 task–model evaluations, only 6.1% are fully solved—underscoring the benchmark's difficulty. The multi-axis rubric nonetheless provides decomposed reward signal even on partially correct runs, making it suitable for reinforcement learning.

Figure 1 plots performance against output cost estimates ($USD per 1M tokens). Opus 4.6 is the most expensive model and also the highest-scoring frontier model. Gemini 3 Flash offers the lowest cost but substantially underperforms, suggesting that cost savings from smaller models do not translate to viable enterprise performance on $C^4$.

### 4.1.2 Performance by $C^4$ Axis

Figure 8 disaggregates scores by the four $C^4$ evaluation axes. Correctness is the primary discriminator (16.6 pp spread between Ebla and Gemini 3 Flash), driven by differences in retrieval accuracy and visual extraction. Composition is the most uniform axis—all models produce well-structured output. Citations varies substantially, with weaker models providing vague references while the strongest cite exact document IDs and page numbers.

**Failure modes.** We identify several failure modes that recur across all frontier models tested. The two most consequential are failures of abstention and visual misinterpretation.

**Failure to abstain.** When a task requires information that does not exist in the corpus, the correct behavior is to say so. Instead, every model we tested generates a plausible-sounding answer—citing documents that exist but do not contain the claimed information. The output is indistinguishable from a well-sourced response without manually checking every citation. This is the most dangerous failure mode in enterprise deployment: a confident wrong answer is worse than no answer, and in domains like compliance or legal, "I don't know" is often the only safe response.

**Visual misinterpretation.** Models can retrieve documents containing diagrams and visual content, but consistently fail to accurately interpret them. One of our environments, Korvo, simulates a chemical company whose corpus includes a safety evacuation diagram embedded on page 15 of a facilities document. Models find the document and attempt to describe the diagram, but misread spatial relationships, omit critical elements, and produce descriptions that would be unsafe to act on. In industries where visual documents encode safety-critical information—evacuation routes, chemical handling procedures, equipment schematics—this is not an edge case but a deployment blocker.

Beyond these, we confirmed three additional patterns:

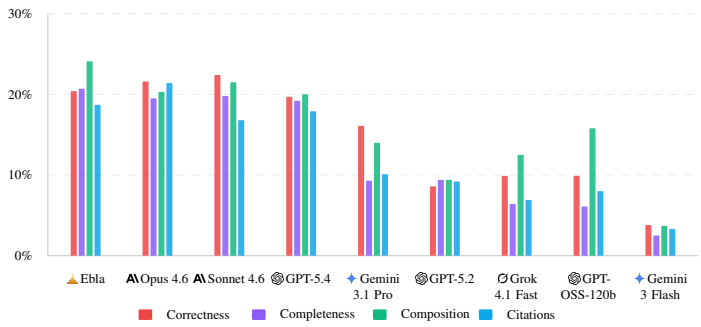



**Figure 8.** $C^4$ evaluation axis scores (%). Models sorted by average across axes.

1. **Cross-document arithmetic fabrication.** When a task requires combining values from multiple documents, every frontier model occasionally invents intermediate numbers rather than retrieving them—then performs arithmetic on fabricated inputs and presents the result with full confidence. Across our evaluation, 23% of tasks involving cross-document arithmetic triggered at least one fabricated intermediate value, and in no case did the model flag uncertainty about the invented number.
2. **Authority hallucination under ambiguity.** When two documents contain contradictory information, models fabricate a resolution—inventing governance hierarchies, override rules, or effective dates that appear in neither source—rather than flagging the conflict. In an enterprise where policy conflicts have legal or compliance implications, this means an agent will silently paper over exactly the kind of discrepancy a human analyst is hired to catch.
3. **Citation drift on long documents.** Models cite specific page numbers that are close to but not the actual source page, particularly on documents exceeding 15 pages. Citations drift toward section headers and table-of-contents pages, eroding the entire point of requiring citations.

#### 4.1.3 Analysis by Document Category

Table 2 disaggregates scores across six document categories and three modalities. Governance, Risk & Compliance yields the highest scores, likely because compliance documents tend to be well-structured. Product, Engineering & Data is the weakest, requiring agents to navigate ambiguous project names and technical diagrams.

#### 4.1.4 Completeness–Correctness Divergence

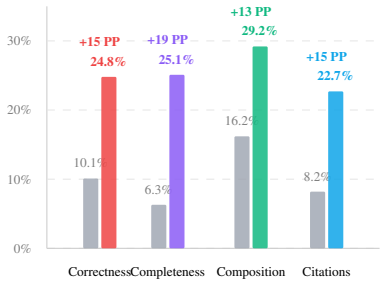
A diagnostic pattern emerges from the axis heatmap: models that locate one source but miss the second corroborating document score well on Correctness but poorly on Completeness. This divergence is most pronounced for Gemini 3.1 Pro (16.1% Correctness vs. 9.3% Completeness) and suggests that multi-hop retrieval—not single-document extraction—is the binding constraint. Citation accuracy correlates with but does not reduce to Correctness: an agent can extract the correct value but cite the wrong page, or cite correctly but misinterpret the evidence.

### 4.2 Training

Using the training setup described in Section 3.5, we evaluate the post-trained model (Ebla) on the full 40-task $C^4$ benchmark.

The post-trained model, Ebla, scores 25.4% mean on the full 40-task benchmark—a +18.3 pp gain over the baseline, 5 pp above the best frontier model (Opus 4.6 at 20.1%).

**Figure 9.** Pass@1 before/after RL training by $C^4$ axis.




Completeness shows the largest absolute gain (+18.8 pp) and the largest relative improvement (4.0× baseline), as the model learned to decompose questions and retrieve evidence for all parts. Correctness gains +14.7 pp (2.5×) as RL directly optimizes against fabrication penalties. Citations gains +14.5 pp (2.8×), with the model learning to ground claims in specific document pages. Composition gains the least (+13.0 pp, 1.8×), consistent with the base model's already higher starting point on this axis. All evaluation runs used a fixed step budget of 36 tool-use calls with no test-time compute scaling (e.g., parallel rollouts or majority voting); reported scores reflect single-pass inference.

### 4.3 Behavioral Changes

Beyond aggregate scores, RL training produced qualitative shifts in how Ebla interacts with the search environment compared to the GPT-OSS-120b baseline.

**Search efficiency.** The baseline model exhibits a pattern common to weaker agents: it issues broad, repetitive search queries that retrieve the same documents multiple times, frequently exhausting its full 36-step budget without converging on an answer. Ebla's trajectories are markedly shorter. Across all 40 evaluation tasks, Ebla uses a median of 14

**Table 2.** Scores (%) by document category and modality. Models sorted by overall performance. Blue intensity encodes score magnitude.

| Model | Document Category | | | | | | Modality | | |
|---|---|---|---|---|---|---|---|---|---|
| | GRC | People | Finance | Sales | Engg | Ops | Text | Tab | Vis |
| Ebla | 7.5 | 27.8 | 32.6 | 22.6 | 38.7 | 19.5 | 29.2 | 11.7 | 24.2 |
| Opus 4.6 | 0.0 | 15.1 | 36.5 | 0.0 | 54.0 | 0.0 | 20.5 | 13.6 | 25.9 |
| Sonnet 4.6 | 7.4 | 33.6 | 23.4 | 9.7 | 28.0 | 17.3 | 24.8 | 0.0 | 16.7 |
| GPT-5.4 | 1.2 | 3.0 | 30.2 | 18.5 | 26.7 | 14.5 | 22.8 | 1.8 | 13.8 |
| Gemini 3.1 Pro | 0.0 | 20.4 | 23.7 | 0.0 | 9.9 | 20.0 | 15.5 | 0.0 | 11.5 |
| GPT-5.2 | 0.0 | 2.2 | 18.3 | 2.6 | 34.8 | 0.0 | 12.7 | 15.8 | 0.0 |
| Grok 4.1 Fast | 0.0 | 0.0 | 14.2 | 14.4 | 12.1 | 0.0 | 12.1 | 0.0 | 0.0 |
| GPT-OSS-120b | 0.0 | 0.0 | 14.2 | 7.4 | 12.8 | 0.0 | 10.1 | 0.0 | 1.9 |
| Gemini 3 Flash | 2.3 | 0.0 | 15.8 | 9.2 | 0.0 | 2.4 | 6.9 | 9.2 | 0.0 |

tool calls per run compared to 31 for the baseline—a 55% reduction—while achieving 3.6× higher scores. On tasks where both models arrive at the correct answer, Ebla reaches it in roughly half the steps, indicating that RL training has improved not just answer quality but search strategy.

**Query diversity and retrieval breadth.** The baseline tends to rephrase the same query with minor variations, retrieving overlapping document sets across successive search calls. Ebla exhibits greater query diversity: it decomposes multi-part questions into distinct sub-queries targeting different document categories, then fetches specific documents to verify claims before committing to an answer. This decomposition behavior emerges naturally from RL optimization—the rubric rewards Completeness independently of Correctness, so an agent that retrieves evidence for *all* parts of a question earns higher reward than one that answers only the first part correctly.

**Calibrated commitment.** Perhaps the most striking behavioral change is Ebla's ability to commit to an answer under partial evidence. The baseline model either fabricates confidently (earning negative Correctness scores) or refuses to answer at all. Ebla occupies a middle ground: when evidence for one sub-question is unavailable, it answers the parts it can verify and explicitly states what it could not find, earning partial Completeness credit without triggering fabrication penalties. This behavior was not explicitly trained—it emerged from the interaction between the multi-axis rubric and the penalty structure.

**Citation precision.** The baseline produces vague document references ("according to the financial report") or cites documents by title without page numbers. After RL training, Ebla consistently provides document IDs and page-level citations, with citation accuracy improving from 8.2% to 22.7% on the Citations axis. On tasks involving documents exceeding 15 pages, the baseline's citations drift toward early pages and section headers; Ebla's citations target the specific pages containing the referenced data, suggesting improved document comprehension during the isolated fetch analysis step.

Interactive task samples comparing baseline and post-trained outputs are available at `https://aviro.ai/`.

## 5 Conclusion

We presented $C^4$, a benchmark for grounded enterprise reasoning that evaluates agents across four independently scored axes using fine-grained rubrics with fabrication penalties. Across 840 runs on seven frontier models, the best model achieves 20.1% and only 6.1% of task–model pairs are fully solved—demonstrating that the combination of retrieval, multi-hop reasoning, visual understanding, and epistemic calibration required by real enterprise search remains a significant open challenge.

After RL training on 30 tasks, Ebla achieves 25.4% on the full 40-task $C^4$ benchmark—surpassing all frontier models. The environments and rubric structure that underpin $C^4$ are not only a rigorous evaluation but a productive training environment.

The benchmark's design ensures that progress is measurable at multiple granularities: the four $C^4$ axes decompose performance into interpretable dimensions, and the signed-weight criteria produce near-continuous reward distributions that provide dense signal for both evaluation and reinforcement learning. The closed, knowledge-graph-generated environments eliminate contamination risk and ensure reproducibility.

Evaluation access is available to research teams and enterprises. For API access, dataset licensing, or RL partnerships, contact `founders@aviro.ai`.

## References

Salesforce Research. HERB: Heterogeneous Enterprise RAG Benchmark. *arXiv:2506.23139*, 2025.

Wix Research. WixQA: Enterprise RAG Benchmark. *arXiv:2505.08643*, 2025.

Databricks. OfficeQA: A Benchmark for End-to-End Grounded Reasoning. Databricks Blog, 2025.

IBM Research. MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. *arXiv:2501.03468*, 2025.

Salesforce Research. UAEval4RAG: Unanswerability Evaluation for RAG. *arXiv:2412.12300*, 2024.

Databricks AI Research. KARL: Knowledge Agents via Reinforcement Learning. Databricks Technical Report, 2026.

OpenAI. ChatGPT Deep Research. https://platform.openai.com/docs/guides/deep-research, 2024.

OpenAI. Reinforcement Fine-Tuning (RFT). https://platform.openai.com/docs/guides/reinforcement-fine-tuning, 2024.

# A WEBSITE SAMPLES

This appendix mirrors the sample section published at https://aviro.ai/c4-ebla#samples. We include the same before/after walkthrough task and eight representative sample tasks used in the website viewer, reformatted for static reading. Tasks are organized by environment: Korvo (chemical manufacturing, 3 tasks), Vexel (financial services, 3 tasks), and Zyro (SaaS analytics, 2 tasks).

## A.1 Before/After Walkthrough: Travel, Audit & Interest Traps

**Task.** A contract review note for AP and Finance evaluating four compliance items against Korvo's master services agreement: a $18,500 travel/housing claim for a 150-day site stay, a $2,200 late-payment interest charge after 3 late payments, a $35,000 auditor fee deduction, and a competitor-status determination for "Nexus Advisory" via Exhibit 31.

**Key sources.**

- Exhibit 4, Sections 7.2(c) and 7.2(d) (travel/housing thresholds).
- MSA Sections 9.4, 15.1, 15.5(b) (interest, audit caps).
- Exhibit 31 — Korvo Global Competitor Registry (absent from corpus).

**Table 3.** Same before/after walkthrough shown on the website, rendered here as a static summary.

| Model | Score | Retrieval behavior | Outcome |
|---|---|---|---|
| GPT-OSS-120b | 38% | Searches broadly for MSA sections, fetches the housing exhibit and audit clause but misses the interest threshold nuance. | Catches the audit cap and travel cutoff but applies the general interest rule instead of the override clause requiring four failures; misses that Exhibit 31 is absent. |
| Ebla | 62% | Targeted searches for each of the four contract provisions, fetches Exhibit 4 and MSA Sections 9.4 and 15.1 before answering. | Identifies all four contract traps—120-day travel cutoff, four-failure interest threshold, $25K audit cap, missing Exhibit 31—and correctly reverses the internal team's position on each item. |

## A.2 Representative Task Set

### A.2.1 Korvo: Travel, Audit & Interest Traps

**Prompt.** Review a contract note evaluating four compliance items against Korvo's MSA: a $18,500 travel/housing claim for a 150-day site stay citing Exhibit 4 Section 7.2(c), a $2,200 late-payment interest charge after 3 late payments citing MSA 9.4, a $35,000 auditor fee deduction citing MSA 15.1, and a competitor-status determination via Exhibit 31. For each item, the agent must quote the controlling clause verbatim and produce a pay/dispute summary table.

**Primary sources.**

- Exhibit 4, Sections 7.2(c) and 7.2(d).
- MSA Sections 9.4, 15.1, 15.5(b); Exhibit 28.
- Exhibit 31 — Korvo Global Competitor Registry (absent from corpus).

**Rubric highlights.**

- The 150-day stay exceeds the 120-day threshold in Section 7.2(d), shifting costs to Provider—must dispute the $18,500.
- Three late payments do not meet the four-failure threshold in MSA 9.4—must dispute the $2,200.
- Audit reimbursement is capped at $25,000 despite the 13% overcharge trigger—must limit deduction.
- Exhibit 31 is absent; fabricating Nexus Advisory's competitor status incurs a $-100$ penalty.

| | | | | | |
|---|---|---|---|---|---|
| Ebla | 62% | GPT-OSS-120b | 38% | Grok 4.1 Fast | 38% |
| Sonnet 4.6 | 30% | Opus 4.6 | 0% | GPT-5.2 | 0% |
| GPT-5.4 | 0% | Gemini 3.1 Pro | 0% | Gemini 3 Flash | 0% |

### A.2.2 Korvo: Housing & Interest Audit Inversions

**Prompt.** Structurally parallel to the previous task but with inverted contract parameters: a $18,500 corporate housing claim for a 7-month Chicago assignment citing Exhibit 4 Section 6.3(f), a $2,800 late-payment penalty after 2 late payments citing MSA 11.5, a $35,000 audit fee deduction on an $800K engagement citing MSA 13.2, and a competitor-status determination for "Triton Global" via Exhibit 22.

**Primary sources.**

- Exhibit 4, Sections 6.3(f) and 6.3(g).
- MSA Sections 11.5, 13.2, 13.6(d); Exhibit 19.
- Exhibit 22 — Korvo Competitor Registry (absent from corpus).

**Rubric highlights.**
- The 7-month stay exceeds the 6-month threshold in Section 6.3(g), shifting housing to Provider—must dispute.
- Two late payments do not meet the three-failure threshold in MSA 11.5—must dispute.
- Audit reimbursement is capped at $14,000 despite the 12% overcharge—must limit deduction.
- Exhibit 22 is absent; fabricating Triton Global's status is severely penalized.

| | | | | | |
|---|---|---|---|---|---|
| GPT-5.4 | 84% | Ebla | 80% | Opus 4.6 | 80% |
| Sonnet 4.6 | 61% | Gemini 3.1 Pro | 35% | GPT-OSS-120b | 17% |
| GPT-5.2 | 0% | Gemini 3 Flash | 0% | Grok 4.1 Fast | 0% |

### A.2.3 Korvo: ECA Caps & Transit Exceptions

**Prompt.** Evaluate Korvo's annual Economic Cost Adjustment (ECA) invoice covering three items: a 4.0% ECA increase to Monthly Base Fees (actual benchmark shift 4.8%, 3-year mean 3.6%), 20 hours of simulation overage billed at $416/hr (standard $400/hr + 4.0% ECA), and $15,000 travel/housing for a 7-month engineer assignment. The agent must produce a compact pay/dispute table with exact dollar amounts.

**Primary sources.**
- Exhibit 4, Section 13.5(k) (3-year mean override of Section 13.2 cap).
- Exhibit 4, Section 13.1(a) (ECA scope: fixed Base Fees only).
- Exhibit 4, Sections 6.3(f) and 6.3(g) (travel/housing threshold).
- Attachment 2.1-A, page 18 (Emergency Support rate schedule).

**Rubric highlights.**
- Section 13.5(k) overrides the Section 13.2 cap: fee growth limited to the lower of the actual shift (4.8%) or the 3-year mean (3.6%)—correct ECA is 3.6%, not 4.0%.
- ECA applies only to fixed Base Fees per Section 13.1(a)—Emergency Support rate cannot have ECA applied; must dispute $320.
- The 7-month stay exceeds 6 months, shifting all travel/housing to Provider—must dispute $15,000.
- Final table: pay $8,000, dispute $15,320.

| | | | | | |
|---|---|---|---|---|---|
| Sonnet 4.6 | 86% | Gemini 3.1 Pro | 85% | Gemini 3 Flash | 51% |
| GPT-5.4 | 48% | Ebla | 34% | Grok 4.1 Fast | 23% |
| Opus 4.6 | 0% | GPT-5.2 | 0% | GPT-OSS-120b | 0% |

### A.2.4 Vexel: Roadmap Deliverable Boundary

**Prompt.** Confirm three roadmap details for a PMO sync: which Product Area owns "External Market Data API Integration" and its original target release date, the sibling deliverable in that same area with the same release date, and three verbatim strategic objectives defining the Vexel Platform Development Strategy.

**Primary sources.**
- *Quarterly Deliverables Report Template*, page 1.
- *Q1 2024 Platform Roadmap & Product Area Planning*, page 4.

**Rubric highlights.**
- "External Market Data API Integration" belongs to the Growth Product Area with target release June 30, 2024.
- The sibling deliverable is "Vexel Pathways 3.0 Skills Inference Engine."
- Three verbatim strategic objectives must contain specific key phrases (e.g., "Empower Product Area teams with autonomous decision-making").
- Misidentifying the Product Area or quoting objectives from a different strategy section is severely penalized.

| | | | | | |
|---|---|---|---|---|---|
| GPT-5.2 | 100% | Ebla | 95% | Opus 4.6 | 95% |
| GPT-5.4 | 51% | Sonnet 4.6 | 0% | Gemini 3.1 Pro | 0% |
| Gemini 3 Flash | 0% | GPT-OSS-120b | 0% | Grok 4.1 Fast | 0% |

### A.2.5 Vexel: Signals Workstream Correction

**Prompt.** A planning crosswalk claims "Vexel Signals v3.2 belongs in the Growth June 30 release cohort with Vexel Pathways 3.0 Skills Inference Engine, and the controlling workstream is FA-06 Vexel Scholars Program Integration." The agent must verify against official materials and return six corrected fields including the correct product area, release date, peer deliverable, and FA-03 focus-area text.

**Primary sources.**
- *Quarterly Deliverables Report Template*, page 1.
- *Product Area Strategy & Workstream Overview*, page 8.

**Rubric highlights.**
- Crosswalk does not stand: Signals v3.2 belongs to the Intelligence Product Area (not Growth), target March 15, 2024.
- Correct peer deliverable is "Sentiment classification v2.8 deployment."
- FA-03 focus-area text is "Engagement Trend Predictive Analytics."
- Must return null for release stage (materials do not provide one); inventing a stage is severely penalized.

| | | | | | |
|---|---|---|---|---|---|
| Sonnet 4.6 | 100% | Opus 4.6 | 83% | Ebla | 70% |
| GPT-5.4 | 55% | Grok 4.1 Fast | 28% | GPT-OSS-120b | 26% |
| GPT-5.2 | 15% | Gemini 3.1 Pro | 15% | Gemini 3 Flash | 15% |

### A.2.6 Vexel: Signals Status Crosswalk

**Prompt.** Extended version of the Signals correction task. The crosswalk now also claims the Growth bundle status note "In Beta with Meridian Consulting Group," area lead Christopher Wu, and FA-06. The agent must return eight corrected fields including the exact Intelligence March-15 bundle status comment, the correct area lead, and an explicit null for owner team.

**Primary sources.**
- *Quarterly Deliverables Report Template*, page 1.
- *Product Area Strategy & Workstream Overview*, page 8.

**Rubric highlights.**
- Same area/date/peer corrections as the v3 task (Intelligence, March 15, Sentiment classification v2.8).
- Exact bundle status comment is "Deployed to production; Mosaic instance validated."
- Correct area lead is "Heather Young (Lead)" (not Christopher Wu from the Growth bundle).
- Must return null for owner team; converting a person name into a team name is severely penalized.

| | | | | | |
|---|---|---|---|---|---|
| Gemini 3.1 Pro | 85% | Ebla | 65% | Sonnet 4.6 | 45% |
| GPT-5.4 | 30% | Opus 4.6 | 0% | GPT-5.2 | 0% |
| Gemini 3 Flash | 0% | GPT-OSS-120b | 0% | Grok 4.1 Fast | 0% |

### A.2.7 Zyro: Salesforce Knowledge Articles

**Prompt.** List all knowledge article names found in the Salesforce demo, using ellipses (. . . ) for truncated titles as they appear on screen. Identify articles to the best of ability even when names are cut off.

**Primary sources.**
- *Customer Service Job Aids*, pages 3, 5, 6, 12, 22, 42–43.
- *Customer Service Skills Drills*, pages 7–8, 11.

**Rubric highlights.**
- Must identify articles by ID (e.g., "HealthSphere Platform: Sync. . . " ID 000006053, "ShopperIQ: Release dates, con. . . " ID 000010913).
- Must identify articles by author (e.g., "HealthSphere Real-World Evidence Data Integration Gu. . . " by Priya Ramachandran).
- Must use ellipses for truncated titles as explicitly requested; must not invent articles or assign incorrect IDs.
- Requires extraction from multiple document sources across numerous pages.

| | | | | | |
|---|---|---|---|---|---|
| Ebla | 5% | Opus 4.6 | 5% | Gemini 3 Flash | 5% |
| Gemini 3.1 Pro | 3% | Sonnet 4.6 | 0% | GPT-5.2 | 0% |
| GPT-5.4 | 0% | GPT-OSS-120b | 0% | Grok 4.1 Fast | 0% |

### A.2.8 Zyro: Platform Coverage Gap Ranking

**Prompt.** For each of the 11 Zyro platforms, determine presence across five coverage dimensions—training curriculum, chat origin support, strategic priority, knowledge expert coverage, and capital/expense investment classification—calculate gap counts, and rank by risk.

**Primary sources.**
- *Platform Mastery Learning Plan*, page 1.
- *Multi-Platform Analytics Workflows*, page 101.

- *Year-End Review*, pages 5 (org chart) and 15 (strategic priorities).
- *Knowledge Framework*, page 7; *Investment Guide*.

**Rubric highlights.**

- No platforms appear in training curriculum; no platforms have dedicated knowledge expert coverage.
- Investment classification applies only to 4 retail platforms; 7 non-retail platforms get null.
- MediScope has the most gaps (5 = Critical risk).
- Division leader accountability cannot be determined from available documents—must abstain rather than fabricate.

| | | | | | |
|---|---|---|---|---|---|
| GPT-5.4 | 27% | Ebla | 20% | Opus 4.6 | 15% |
| Gemini 3 Flash | 12% | Gemini 3.1 Pro | 6% | Sonnet 4.6 | 0% |
| GPT-5.2 | 0% | GPT-OSS-120b | 0% | Grok 4.1 Fast | 0% |

The live website version additionally exposes full prompts, expanded rubric criteria, and trace links for every model result: https://aviro.ai/c4-ebla#samples.