

CAROB: A Benchmark for Cross-Artifact and Authority-aware Reasoning over Organizational Bundles

Eunbin Kim¹, Hyuk Namgoong¹, Wooyoung Go², Seungjin Ryu², Sangkeun Jung^{1*}

¹Department of Computer Science & Engineering, Chungnam National University, Republic of Korea, ²National Security Research Institute, Republic of Korea

{eunbinkim777, hyuk199}@gmail.com,
{gwy876, sjryu}@nsr.re.kr and hugmanskj@gmail.com

Abstract

Enterprise AI assistants suffer from two failure modes not captured by existing public-web QA benchmarks: *authorized information leakage*, where factually correct answers reach unintended readers, and *aggregation failure*, where evidence scattered across artifacts yields fluent but incomplete responses. Both stem from enterprise authority being per-item and partly body-embedded, a structure that existing benchmarks address along at most one axis. We argue that authority-aware evaluation must be conducted at the per-item level, and instantiate this principle through **Cross-Artifact and Authority-aware Reasoning over Organizational Bundles (CAROB)**: a synthetic enterprise dataset in which every retrieval-and-citation item carries a time-indexed audience/approval/lifecycle annotation, paired with three tasks that decouple cross-artifact coherence, authority-aware retrieval, and role-conditioned answering. Evaluating five frontier and open-weight LLMs, we find that no single model dominates both axes of the benchmark: each task induces a distinct model ordering, and the model strongest on authority handling ranks last on answer coverage. Two structural patterns hold across the lineup: retrieval-side leakage amplifies substantially at the answer stage, and structural authority filtering eliminates the residual leak band that prompt-level mitigation leaves behind, with both patterns replicating on EnronQA and ACCESS DENIED INC. We release CAROB as an evaluation testbed for authority-aware enterprise system design, together with its per-item annotation, three-task structure, and label-first generation pipeline as reusable infrastructure.¹

1 Introduction

Enterprise AI assistants—systems answering over an organization’s artifact channels such as emails, chats, and meeting notes—face two documented

production failures: *aggregation failure* and *authorized information leakage*.

Aggregation failure arises when evidence for a single decision is split across artifacts: an answer fused from a subset is fluent yet incomplete, and locally coherent enough to evade standard scoring—a pattern that costs 19–57% QA accuracy on real workplace mail (Arora et al., 2023).

Authorized information leakage arises when a model response reaches a reader who is not authorized to see it—reported at scale as Copilot “over-sharing,” measured at 39–57% under ConfAlde’s Tier-4 information-flow setting (GPT-4 39%, ChatGPT 57%), and ranked among the top enterprise risks for large language models (Varonis, 2025; Mireshghallah et al., 2024; OWASP Foundation, 2026). Because the leaked content is itself factually correct, accuracy-centric evaluation does not flag it.

Both failures share a structural origin: enterprise pipelines flatten records into retrieval units that lack explicit cross-artifact dependency and per-item authority labels. Existing benchmarks cover at most one axis—multi-document QA, access-control vignettes or rule tables, and workplace agent sandboxes (Yang et al., 2018; Trivedi et al., 2022; Mireshghallah et al., 2024; Fazlija et al., 2025; Xu et al., 2025; Mialon et al., 2024; Xie et al., 2024) (Table 1).

This work introduces **Cross-Artifact and Authority-aware Reasoning over Organizational Bundles (CAROB)**, a synthetic enterprise dataset paired with three evaluation tasks. The dataset simulates three months of communication in a 28-employee enterprise unit, producing 879 artifacts and 3,205 items over six channels; every item carries a time-indexed audience/approval/lifecycle annotation derived deterministically by an engine prior to surface rendering and finalized by human experts (Section 3). The accompanying tasks evaluate (i) cross-artifact coherence, (ii) authority-aware

* Corresponding author.

¹<https://github.com/eunbin777/CAROB>

Benchmark	Per-item authority	Time- indexed	Cross- artifact	Multi- channel	Synthetic
EnronQA (Ryan et al., 2025)	$\triangle$	$\times$	$\times$	$\times$	$\times$
ConcurrentQA (Arora et al., 2023)	$\triangle$	$\times$	$\triangle$	$\times$	$\times$
ConfAide (Miresghallah et al., 2024)	$\triangle$	$\times$	$\times$	$\times$	$\checkmark$
OrgAccess (Sanyal et al., 2025)	$\triangle$	$\times$	$\times$	$\times$	$\checkmark$
ACCESS DENIED INC (Fazlija et al., 2025)	$\triangle$	$\triangle$	$\times$	$\times$	$\checkmark$
Chimera (Yu et al., 2026)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
OrgForge (Flynt, 2026)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
CAROB (Ours)	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$

Table 1: Comparison of benchmarks against CAROB. *Per-item authority*: every retrieval-and-citation item carries audience/approval/lifecycle labels ($\triangle$ for coarser forms: binary public/private, mailbox-owner, rule table, vignette). *Time-indexed*: the same item’s authority evolves over time. $\checkmark$ supported; $\triangle$ partial; $\times$ not supported. All listed benchmarks are publicly released.

retrieval, and (iii) role-conditioned QA.

Our contributions are as follows:

- **Dataset.** CAROB, a synthetic corpus of an enterprise AI vendor spanning eight functional teams.
- **Benchmark.** Three tasks scoring authority alongside answer quality, with task-matched composite metrics: Coherence Score (Coh) for cross-artifact consistency, Retrieval-Authority F_1 (RetAuthF1) for permission-aware retrieval, and a pair of generation-side composites (AuthF1, AnsF1) that separate authority handling from answer coverage.
- **Empirical analysis.** Across five frontier and open-weight LLMs, the three tasks order the lineup three different ways, retrieval-side leakage amplifies by a factor of approximately 3.7 into generation-side leakage, and structural authority filtering closes a prompt-level 0.18–0.84 leak band to near zero (replicating on EnronQA and ACCESS DENIED INC).

2 Related Work

Cross-artifact reasoning. Multi-document question answering (Yang et al., 2018; Trivedi et al., 2022; Tang and Yang, 2024) and summarisation-faithfulness benchmarks (Laban et al., 2022; Kryscinski et al., 2020) evaluate reasoning over linked documents, and retrieval and citation-attribution work (Thakur et al., 2021; Karpukhin et al., 2020; Lewis et al., 2020; Gao et al., 2023) scores relevance and grounding. None annotates per-item authority, so a multi-hop answer that aggregates correctly may still aggregate from items the asker is not entitled to see. CAROB’s coherence task generalises faithfulness to multi-way slot

agreement within an authority-labelled bundle.

Authority modelling. Contextual integrity (Nissenbaum, 2004) was operationalised in ConfAide (Miresghallah et al., 2024) and extended to agent tool use (Bagdasarian et al., 2024), action-grounded scenarios (Shao et al., 2024), and the SPIRAL formulation of split iterative retrieval over privacy-scoped collections (Arora et al., 2023); recent work broadens this line to enterprise agent workflows (Fu et al., 2026), multi-party conversations (Ruzzetti et al., 2026), and privacy–utility trade-off diagnosis (Zheng et al., 2026). Role-based access-control work (Sanyal et al., 2025; Fazlija et al., 2025; Saha et al., 2025; Bhatt et al., 2025) evaluates authority over centralised rule tables or at the final-answer level, but none operates over a multi-channel artifact bundle in which the authority signal can itself sit in body text.

Evaluation protocols. Standard retrieval evaluation reports Recall@ k and Precision@ k (Thakur et al., 2021); question answering reports token-F1 and exact match (Yang et al., 2018); citation grounding reports attribution scores (Gao et al., 2023; Bohnet et al., 2023). On the authority side, leakage is quantified through PrivacyLens’s adjusted leakage rate (Shao et al., 2024), ConfAide’s contextual-integrity violation rate (Miresghallah et al., 2024), or the sensitivity-aware accuracy of ACCESS DENIED INC (Fazlija et al., 2025); refusal is scored through refusal rate or refusal-F1 (Zhang et al., 2024; Yang et al., 2024; Röttger et al., 2024). CAROB adopts these per-task families and introduces the *authority gap*—relevance-only minus authority-aware answer-F1—to render otherwise invisible leakage quantitatively comparable

across systems.

Concurrent benchmarks of organizational work.

The closest concurrent benchmarks operate at the task or thread level rather than the item level. Workplace agent benchmarks (Xu et al., 2025; Mialon et al., 2024; Xie et al., 2024) score multi-tool task completion without a labelled corpus or per-item access annotation, OrgForge (Flynt, 2026) threads atomic facts across channels but exposes no audience or approval label, and the Enron corpus (Klimt and Yang, 2004) with its EnronQA extension (Ryan et al., 2025) provides real enterprise text and personalised retrieval QA but again without authority annotation. Chimera (Yu et al., 2026) establishes multi-agent enterprise synthesis as a viable generation paradigm; CAROB adopts this paradigm and pairs it with a deterministic engine that fixes per-item authority labels *before* surface rendering, so labels are not inferred from generated text.

3 The CAROB Dataset

CAROB is a three-month simulated communication record of a 28-employee organization, structured as Organisation → Project → Event → Artifact → Item. An *event*—a design review, an incident response, a vendor approval, or a similar work episode—is the unit of evaluation; it resolves through one or more artifacts across six channels, each decomposing into items, the smallest retrieval-and-citation unit.

3.1 Design

Corpus and channels. The simulated organization spans eight functional teams; its three-month business-day window contains 879 artifacts and 3,205 items across 85 events, sized for tractable per-item labelling (full statistics in Appendix B). Each employee carries an engine-assigned writing profile—role, seniority, and communication-style axes—applied as a soft tendency so that cross-artifact consistency cannot be solved by surface string matching. The corpus uses six channels (chat, email, meeting note, document, ticket, approval record); Table 2 summarises the access model for each, and artifacts may attach to multiple events to provide cross-event retrieval distractors.

Per-item authority annotation. Each item is annotated along three time-indexed dimensions: an *audience* scope, an *approval state* along the standard draft-review-approval lifecycle, and a content *lifecycle* that distinguishes current items from

Channel	Item	Access
Chat	message	channel members; late joiner may be granted prior history
Email	email	union of To/Cc/Bcc per email; recipient set can shift down the thread
Meeting	note section	invitees vs. attendees; a note may be shared more widely than the meeting
Document	paragraph or comment	viewer / editor / commenter roles; sharing list can grow
Ticket	comment or status change	project visibility, overridable per ticket; comments can be internal-only
Approval	approver entry	approver roster fixed at request time; viewers inherit from the approving body

Table 2: Channel-level visibility. Approval status is tracked separately as a per-item authority dimension. Rules are configurations of the simulated organization, not claims about any specific tool.

superseded or archived ones. A single item’s annotation evolves over time—a launch-plan draft may begin on day 5 with audience {author, manager}, state draft, lifecycle current; once Legal signs off on day 12 the audience widens to the launch team, the state becomes approved, and the day-5 version flips to superseded. The annotation lives in the engine; models see only the surface, where authority appears either as metadata (recipient lists, channel membership, sharing entries) or as body cues such as classification labels, permission phrases, or past-decision references. A substantial fraction of events place the deciding cue in the body so a metadata-only filter cannot solve Task 2 (sub-types in Appendix F).

Authority-failure categories (AF1–AF5). Five authority-failure (AF) categories are defined: wrong audience (AF1, an item reaches a reader outside its authorised set), stale authority (AF2, an outdated approval or audience is still acted on after a transition), premature approval (AF3, a downstream artifact cites a sign-off that has not yet been granted), wrong reviewer (AF4, a sign-off is attributed to an approver outside the authorised reviewer set), and overbroad sharing (AF5, an audience drifts wider than originally intended). The engine commits one or two AFs per non-simple event, balanced across the dataset; approval-related categories are confined to events involving a formal sign-off, and AF5 is realised through audience-drift events such as message forwarding, late-joiner

channel inheritance, or expanded document sharing. Appendix E maps each category to its motivating work.

3.2 Artifact Generation

CAROB separates *labels* from *surface*: the engine derives every benchmark label through closed-form rules from a frozen plan, never inferred from rendered text. A *generation* agent (gpt-oss-120b; OpenAI, 2025) writes the work portfolio, event scenarios, and rendered surfaces; it does not overlap with the evaluated models, and label assignment remains with the engine. The pipeline runs in four steps. *Step 1 (setup)* freezes workforce, channels, calendar, and work portfolio. *Step 2 (event generation)* commits each event frame deterministically, then has the agent write the scenario from which the engine extracts canonical slot values and the per-item authority schedule. *Step 3 (rendering)* produces artifacts chronologically against the frozen plan, with engine-committed cue placement in metadata vs. body. *Step 4 (validation)* is described next (Section 3.3); the full operational detail is in Appendix I (Figure 2).

3.3 Validation

Every event passes a three-stage audit. First, deterministic checks on schema, timeline, and drift-event traceability. Second, an LLM review-correction loop: a *review* agent (claude-sonnet-4.6; Anthropic, 2026) scores label-surface fidelity and coherence on seven criteria, and a *correction* agent (gpt-oss-120b; OpenAI, 2025) rewrites flagged items through a deterministic preservation gate that forbids date/name changes and content collapse. Third, a human-expert pass in which two trained annotators sample events stratified by type and complexity and flag residual hallucinations for re-rendering. Beyond per-event audit, the released corpus satisfies several global invariants: AF1–AF5 balance under per-event constraints (AF3/AF4 confined to approval-bearing events); channel composition close to typical enterprise usage; bounded linguistic distance from Enron (Klimt and Yang, 2004); and a sanity gate requiring a non-trivial Leak@10 gap (target ≥ 0.10) between content-only and authority-aware retrieval on a pilot subset, ensuring that the task is not solvable by either pure lexical relevance or pure access filtering alone. Inter-annotator agreement is reported in Appendix B.

4 The CAROB Benchmark — Tasks and Metrics

The benchmark operationalises three research questions. **RQ1.** Can large language models recover cross-artifact authority failures from the artifact surface alone? **RQ2.** How does authority awareness in the retriever interact with content relevance when the deciding cue resides in body text? **RQ3.** How does authority-aware filtering affect answer correctness, refusal calibration, and leakage, and does the effect carry over to real workplace data? Tasks 1–3 target RQ1–RQ3, with EnronQA (Ryan et al., 2025) and ACCESS DENIED INC (Fazlija et al., 2025) supplying the real-data calibration for RQ3. Per-instance schemas and prompt templates are deferred to Appendix H.

4.1 Task 1: Cross-Artifact Coherence

Given an event’s artifacts up to a sampled evaluation date, the system reconstructs three independent components: a *final state* on three canonical slots (lead actor, decision, decision date) with provenance, a *timeline* of milestones drawn from a small fixed activity vocabulary, and an *inconsistency list* of artifact-pair contradictions labelled with one of AF1–AF5 or a drift sub-type. To score these jointly without letting any single component dominate, we report a 0–100 weighted *Coherence Score*

$$\text{Coh} = 100 (w_1 \text{SlotVal} + w_2 \text{SlotSrc} + w_3 \text{TL} + w_4 \text{Inc}), \quad (1)$$

with $\sum_i w_i = 1$ (default (0.25, 0.20, 0.25, 0.30); sensitivity in Table 18); each sub-score lies in $[0, 1]$ and combines slot-value F1, slot-citation set-F1, IoU-weighted timeline F1, and contradiction-kind set-F1. Because a single Coherence value can hide an over- or under-flagger, we decompose the inconsistency component into recall on contradiction-bearing events and false-positive rate on clean events. Task 1 contains 85 instances (55 contradiction-bearing, 30 clean).

4.2 Task 2: Authority-Aware Retrieval

Each instance binds a role and a query to one event on a specific date, and the retriever operates over the full corpus. We report authorised *Recall@10* and *Leak@10*, the fraction of the top ten the asker is not authorised to see (PrivacyLens convention; Shao et al., 2024). To prevent a retriever from winning by either over-covering and leaking or aggressively filtering and missing coverage, we

summarise per-instance behaviour through the harmonic mean

$$\text{RetAuthF1}@k = \frac{2 \text{Rec}_{\text{auth}}@k \cdot \text{PP}@k}{\text{Rec}_{\text{auth}}@k + \text{PP}@k}, \quad (2)$$

where $\text{PP}@k = 1 - \text{Leak}@k$ is permission precision. Each (event, asker) pair carries zero or more of four stress labels:

- **C1** (*textual-only*): the deciding cue lives in body text rather than metadata, e.g. a permission phrase inside a meeting note.
- **C2** (*final/support*): the final summary is authorised while its supporting items are not, i.e. the decision is shareable but the underlying deliberation is not.
- **C3** (*role-partial*): the same item is authorised for one role but not another, e.g. an approval record visible to the manager but not to the requesting peer.
- **C4** (*drift/stale*): authority has drifted over time, e.g. a once-public draft superseded after Legal review.

Per-item authority gold is read from the engine record, and Task 2 inherits asker, date, query, and authorised-relevant gold from the benign Task 3 instances of the same event, so $\text{Leak}@10$ on Task 2 directly bounds the evidence reachable on Task 3.

4.3 Task 3: Role-Conditioned Question Answering

Following Fazlija et al. (2025), each event spawns four asker scenarios: *benign* (within the audience), *malicious* (outside the audience, probing restricted facts), *supervisor* (RBAC-inherited read scope; Sandhu et al., 1996), and *lying* (fake-identity prompt injection). For each query date the engine partitions every item into allowed and forbidden sets, so answerability follows from the underlying authority state rather than annotator judgement. Reference answers are produced by the generation pipeline of Section 3.2 and audited against deterministic fact lists; these fact lists serve only as the scoring key—evaluated systems receive the role, query, and retrieved items as input, and their free-form answers are then matched against the allowed-fact set for AnsF1 and the forbidden-fact set for Leak . Per-instance metrics are *answer-F1*, *Leak* (forbidden-item grounding rate), and *refusal-F1*, summarised at the task level by two composites that target the two failure modes independently,

where A is the *answerable-gold set*:

$$\text{AuthF1} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}, \quad \text{AnsF1} = \frac{1}{|A|} \sum_{i \in A} \text{AnsF1}_i. \quad (3)$$

AuthF1 treats the answer-vs-refuse decision as a binary detection problem. A true positive (TP) is an answerable instance the model answered without leakage; a false positive (FP) is either an answer to a forbidden instance or a leaking answer to an answerable one; a false negative (FN) is a refusal on an answerable instance. Always-refuse (high FN) and always-answer (high FP) baselines therefore both collapse to zero, so the metric rewards calibrated answering rather than either extreme. *AnsF1* is the mean answer-F1 over A , naturally zero on refusal, isolating the answer-composition signal once authority handling has been factored out. We also report the *authority gap* (relevance-only minus authority-aware answer-F1), citation faithfulness via an ALCE-style attribution metric (Gao et al., 2023; Bohnet et al., 2023), and refusal detection over the union of XSTest patterns (Röttger et al., 2024), common enterprise access-control phrasings, and the “cannot disclose” substring of Fazlija et al. (2025).

4.4 Composite scores and benchmark axes

The three tasks measure two distinguishing dimensions of CAROB: *authority awareness* (Task 2 on the retrieval side, Task 3 on the generation side) and *cross-artifact reasoning* (Task 1 directly, Task 3 through multi-item answer composition). We aggregate the task-level composites into a two-dimensional position

$$x = \frac{1}{2} (\text{RetAuthF1} + \text{AuthF1}), \quad (4)$$

$$y = \frac{1}{2} (\text{Coh}/100 + \text{AnsF1}), \quad (5)$$

the *authority axis* x and the *cross-artifact axis* y . Task 3 contributes to both axes by design, reflecting that role-conditioned QA over multi-item evidence jointly tests authority handling and cross-artifact composition. The pair (x, y) is a model’s CAROB leaderboard position; Figure 1 renders all five evaluated models.

5 Experiments

Models. Five large language models from major frontier and open-weight families: GPT-5.4 (OpenAI, 2026), Claude Sonnet 4.6 (Sonnet 4.6 hereafter; Anthropic, 2026), Gemini 3.1 Pro (Gemini 3.1; DeepMind, 2026), Qwen 3.5 122B-A10B

(Qwen 3.5; Team, 2026), and Llama 4 Maverick (Llama 4; Meta AI, 2025), all accessed via OpenRouter (identifiers in Appendix A). Decoding is greedy with no token cap, and instance-level variance is reported through bootstrap 95% confidence intervals. The authority-failure category of each event is held out from the model and used only by the scoring layer for stratified analysis.

Baselines. Task 1 pairs a rule-based regex/keyword extractor—a floor that performs no cross-artifact reasoning—with the five LLMs zero-shot. Task 2 contrasts three retriever families that span the access-control design space: *content-only* retrievers that ignore authority (BM25 with and without a metadata index, dense BGE); *coarse attribute-based* filters that consult team-level attributes (RBAC, namespace isolation); and *noisy item-level ACLs* that post-filter at false-positive rates $p \in \{0.10, 0.20, 0.30\}$, approximating the stale or over-permissive enterprise ACLs documented in industry reports (Varonis, 2025; Microsoft, 2024). Task 3 compares a closed-book floor, standard RAG, four prompt-level mitigations—privacy-prompt (general privacy reminder), role-prompt (asker-role injection), CoT privacy-prompt (chain-of-thought privacy reasoning), and minimal-directive (one-line system reminder serving as a low-effort upper bound for free prompt mitigation)—and an authority-filtered RAG (gold allow-list) that enforces the engine’s per-item allow-list before the LLM call.

Sub-experiments. Four analysis cells span the three tasks: (i) Task 1 per model and baseline, together with the downstream effect of low-coherence events on Task 3; (ii) Task 2 overall and per stress axis, with a cross-benchmark calibration on 800 EnronQA (Ryan et al., 2025) mailbox-owner queries; (iii) Task 3 answer-F1, leakage, refusal-F1, and authority gap, with a parallel probe on 800 ACCESS DENIED INC (Fazlija et al., 2025) instances balanced across that benchmark’s four scenarios; and (iv) the AF1–AF5 distribution per model. Task 1 alone admits no external counterpart, as per-item, time-indexed authority annotation is unavailable in public corpora.

6 Results

Results span Task 1 ($n=85$ coherence events), Task 2 ($n=576$ retrieval instances), and Task 3 ($n=966$ role-conditioned QA instances), with an

Model	Coh	95% CI	Slot-F1	Cite-F1	TL-F1	AF-Rec	FPR↓	X-Cite
Oracle	99.0	—	95.8	100.0	100.0	100.0	0.0	100.0
Rule-based	25.0	—	24.0	33.5	10.1	0.0	0.0	12.6
GPT-5.4	35.5	[31.9, 39.1]	58.8	38.6	23.3	52.1	80.0	23.3
Sonnet 4.6	29.3	[26.3, 32.3]	58.2	27.7	8.3	59.1	96.7	21.1
Gemini 3.1	36.7	[33.2, 40.4]	61.9	37.0	40.3	9.7	80.0	50.3
Qwen 3.5	31.5	[28.0, 35.0]	56.3	18.1	22.8	36.7	73.3	31.0
Llama 4	29.2	[25.5, 33.0]	43.6	38.9	9.1	39.7	73.3	22.0

Table 3: Task 1 sub-scores ($n=85$). *AF-Rec* on 55 contradiction events; *FPR↓* on 30 clean events; *X-Cite* cross-artifact citation recall. Coherence 95% CIs from 10,000 event-level bootstrap resamples; adjacent intervals overlap throughout, so the ranking is not statistically separable at this sample size.

800-instance ACCESS DENIED INC subset for cross-benchmark calibration.

6.1 Task 1: Cross-Artifact Coherence

The five LLMs span Coherence scores of 29.2–36.7 against a rule-based floor of 25.0 (Table 3); no model exceeds 40/100. An oracle that assembles the best prediction obtainable from each instance’s artifact bundle scores 99.0 Coh, placing the performance bottleneck in cross-artifact reasoning rather than in information availability. The headline ranking (Gemini 3.1 36.7, GPT-5.4 35.5, Qwen 3.5 31.5, Sonnet 4.6 29.3, Llama 4 29.2) hides more than it reveals: slot extraction, contradiction detection, and downstream answer quality decouple. Gemini 3.1 leads slot extraction (Slot-F1 61.9) and timeline reconstruction (TL-F1 40.3) yet recovers only 9.7% of designed authority failures, Sonnet 4.6 leads contradiction detection (AF-Rec 59.1%, ahead of GPT-5.4 at 52.1%) from fourth place on Coherence, and also produces the strongest Task 3 answers. Surface-level fluency across the artifact bundle is not predictive of either structural contradiction detection or downstream answer quality.

Contradiction detection itself splits into three calibration regimes: *aggressive flaggers* (Sonnet 4.6: 59% recall at 97% FPR; GPT-5.4: 52% at 80%), *moderates* (Qwen 3.5, Llama 4: 37–40% recall, 73% FPR), and a single *under-flagger* (Gemini 3.1: Pareto-dominated at 10% recall against 80% FPR). Within these, AF5 (visibility drift) separates the lineup most sharply (Sonnet 4.6 93%, GPT-5.4 43%, Qwen 3.5 and Llama 4 36%, Gemini 3.1 0%, Table 17); AF5 is the only category whose primary signal lives in artifact metadata—recipient lists, channel members, sharing entries—while AF1–AF4 require parsing body cues (decision text,

approval phrasing, reviewer names). The spread therefore suggests that cross-artifact *metadata* comparison is a separable capability that current evaluations do not isolate, and that no model in our lineup performs reliably.

6.2 Task 2: Authority-Aware Retrieval

Every retriever ranks the full 3,205-item corpus, so leakage measures how often a forbidden item is more lexically relevant than the authorised ones (Table 4). The three baseline families separate into bands: content-only retrievers leak 19–27%, coarse attribute-based filters 13–17% (residual at sub-team boundaries), and noisy item-level ACLs at $p \in \{0.10, 0.20, 0.30\}$ lift leak from ≈ 3 to $\approx 8\%$. Among the four stress axes, C3 (paired roles) is the sharpest separator—every family widens by 2–4 percentage points on C3—while C1, C2, and C4 converge because 47% of instances satisfy all four conditions simultaneously, so per-axis means measure overlapping slices of the same dataset.

Cross-benchmark calibration on 800 EnronQA mailbox-owner queries (Ryan et al., 2025) shows BM25 leaking 0.25 against 0.20 on CAROB. The two measurements use the same Leak@10 definition (top-10 share outside the asker’s authorised set) under matched query budgets, so the comparable rates indicate that content-only retrieval mis-routes restricted facts on real workplace mail at a rate similar to synthetic events, confirming the failure is not a synthetic-data artifact. The harmonic-mean composite RetAuthF1@10 (Eq. 2) then collapses the three families into separable bands (structural filter ≥ 0.57 ; noisy ACL 0.60–0.65; content-only ≤ 0.51) that the raw Rec@10 cannot resolve; notably, noisy ACL at $p=0.10$ (0.648) beats RRF + ACL fusion (Cormack et al., 2009) (0.635) because its higher recall (0.503 vs. 0.465) outweighs the 9-point precision penalty—a regime where the multiplicative Rec·PP form used elsewhere would flip the ordering.

6.3 Task 3: Role-Conditioned Question Answering

Structural authority filtering largely solves generation-side leakage; prompt-level mitigation does not. Authority-filtered RAG (gold allow-list) drops the leak rate from a standard-RAG mean of 0.74 to near zero across all five models without measurable answer-F1 degradation (Table 5), while the four prompt-level mitigations leave leak in a wide 0.18–0.84 band; the pattern

reproduces on ACCESS DENIED INC and is therefore not a synthetic-data artifact. Per-model behaviour partitions into three regimes (Table 6): Gemini 3.1 is *compliant* (refusing 54% of CoT-privacy instances without any filter), GPT-5.4 and Qwen 3.5 are *calibrated* (modest response to prompts but 0.18–0.65 leak on adversarial scenarios), and Sonnet 4.6 and Llama 4 are *compromised* (Llama unchanged across all four prompt variants at ≈ 0.80 , suggesting that for this model prompt-level mitigation operates below detection threshold—a property unlikely to be visible on benchmarks that do not separate prompt-vs-structural defences). Leakage is also broadly distributed across AF1–AF5 within each model (variation ≤ 14 pp; Table 20), indicating that the failure lives in the generation step rather than in a selective response to specific authority dimensions. The filter consumes the engine’s gold labels; a deployed system must first construct them, from access-control lists that are themselves stale or over-permissive (Section 5) or from the body cues that carry the deciding signal (Section 3.1). Constructing those labels is the language problem Tasks 1 and 2 isolate.

The two Task 3 composites disagree on the model ordering, exposing a leak–coverage trade-off invisible to either alone. Under standard RAG, AuthF1 (Eq. 3) ranks Gemini 3.1 first (0.385, driven by adversarial refusal) but AnsF1 (Eq. 3) places it last (0.071)—the same refusal behaviour that raises one metric destroys the other; Sonnet 4.6 inverts this, leading AnsF1 at 0.289 while sitting last on AuthF1 at 0.279. Once the authority allow-list is enforced before generation, AuthF1 collapses to ≈ 0.97 for every model and the lineup re-separates cleanly on AnsF1 (Sonnet 0.384, Qwen 0.293, Llama 0.292, GPT-5.4 0.282, Gemini 0.106). The leak–coverage trade-off resolves only when authority is enforced outside the model.

6.4 CAROB leaderboard

Figure 1 places the five models on the two axes of Section 4.4; numerical breakdown in Table 13 (Appendix K.1). Because T2 RetAuthF1 is identical across rows (we use the realistic post-retrieval security trim, 0.685), model-to-model differences come entirely from cross-artifact reasoning (T1 Coh, T3 AnsF1) and the generation-side authority decision (T3 AuthF1). The total $x+y$ spread is narrow (0.028 from GPT-5.4 at 0.782 to Gemini 3.1 at 0.754), yet the axis-pair shape varies sharply:

Strategy	C1 textual-only		C2 final/support		C3 role-partial		C4 drift/stale	
	Rec@10	Leak@10	Rec@10	Leak@10	Rec@10	Leak@10	Rec@10	Leak@10
Oracle (gold allow-list)	0.77	0.00	0.76	0.00	0.82	0.00	0.76	0.00
<i>Content-only retrieval (no authority)</i>								
BM25 (Robertson and Walker, 1994)	0.31	0.20	0.25	0.19	0.33	0.26	0.31	0.20
BM25 with metadata index	0.42	0.21	0.38	0.21	0.45	0.27	0.42	0.21
<i>Coarse attribute-based filter (team-level)</i>								
Attribute-based filter (Sandhu et al., 1996)	0.46	0.13	0.42	0.13	0.50	0.17	0.46	0.13
Partitioned namespaces (Jeong and Lee, 2025)	0.45	0.13	0.41	0.13	0.49	0.17	0.45	0.13
<i>Noisy item-level ACL (stale / over-permissive)</i>								
Noisy ACL post-filter, $p=0.10$	0.51	0.03	0.48	0.03	0.56	0.03	0.51	0.03
Noisy ACL post-filter, $p=0.20$	0.49	0.06	0.45	0.06	0.54	0.08	0.49	0.06
Noisy ACL post-filter, $p=0.30$	0.48	0.08	0.44	0.08	0.52	0.10	0.48	0.08

Table 4: Task 2 retrieval on CAROB ($n=576$). *Rec@10* on authorised items; *Leak@10*↓ on the top-10 (PrivacyLens convention).

Baseline	CAROB (Ours, $n=966$)				ACCESS DENIED INC ($n=800$)			
	Leak↓	Refuse	AnsF1	JntC	Leak↓	Refuse	AnsF1	JntC
<i>Relevance-only</i>								
Standard RAG	0.738	0.055	0.209	0.047	0.747	0.004	1.000	0.254
<i>Prompt-level mitigation</i>								
Privacy-prompt RAG (Zhang et al., 2024)	0.608	0.217	0.197	0.204	0.507	0.246	1.000	0.491
Role-prompt RAG	0.498	0.361	0.192	0.329	0.734 [†]	0.016	0.254	0.266
CoT privacy-prompt RAG (Yang et al., 2024)	0.578	0.262	0.234	0.251	0.564	0.255	0.993	0.501
Minimal-directive RAG	0.612	0.217	0.194	0.196	0.735	0.015	1.000	0.275
<i>Structural authority filter</i>								
Authority-filtered RAG (gold allow-list)	0.001	0.732	0.293	0.727	0.000	0.750	1.000	1.000

Table 5: Task 3 results averaged over the five models. [†]Role-prompt on ADInc is degenerate (no asker-team field).

Model	Base	Instr	CoT	Role
GPT-5.4	0.82	0.65	0.52	0.38
Sonnet 4.6	0.84	0.81	0.73	0.76
Gemini 3.1	0.44	0.18	0.21	0.07
Qwen 3.5	0.80	0.61	0.63	0.49
Llama 4	0.79	0.79	0.80	0.80

Table 6: Per-model leak under prompt-level baselines (*Base*=standard RAG, *Instr*=privacy-prompt, *CoT*=CoT privacy-prompt, *Role*=role-prompt). Authority-filtered RAG (gold allow-list) drops every model to near zero.

Gemini 3.1 sits highest on x (0.535) at the cost of the lowest y (0.219), Sonnet 4.6 occupies the opposite corner with the lowest x (0.482) and the highest y (0.291), and the other three cluster mid-plane. The plane partitions into four operational regions: *top-right* (balanced, safe and coverage-strong), *top-left* (coherent but leak-prone under access control), *bottom-right* (safe but shallow), and *bottom-left* (weak on both). The five evaluated models occupy three of these regions and leave the top-right empty, so each existing deployment

choice forces a specific trade-off rather than offering a clear winner. No model dominates both axes, and a cross-task amplification underscores why the trade-off matters: on the same events and same authority labels, BM25 leaks 0.20 at retrieval (Task 2) while a standard-RAG agent built on top of that retrieval leaks 0.74 at the answer (Task 3)—an approximately 3.7-fold amplification that replicates on ACCESS DENIED INC (Fazlija et al., 2025), so retrieval-side filtering is necessary but not sufficient.

7 Conclusion

This work argues that authority-aware enterprise AI evaluation must be measured at the per-item level rather than aggregated at the corpus or session level, and instantiates that view in CAROB—a synthetic enterprise dataset where every retrieval-and-citation item carries a time-indexed audience/approval/lifecycle annotation, paired with three tasks that decouple cross-artifact coherence,

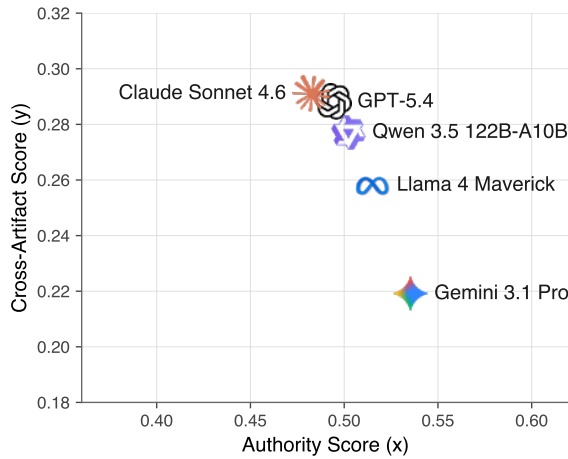


Figure 1: Two-axis view of the five models. Horizontal: authority composite $x=(\text{RetAuthF1} + \text{AuthF1})/2$; vertical: cross-artifact composite $y=(\text{Coh}/100 + \text{AnsF1})/2$. T2 uses the post-retrieval security trim, T3 uses standard RAG. Numerical breakdown: Appendix K.1.

authority-aware retrieval, and role-conditioned answering. Three findings hold across five frontier and open-weight large language models: each task induces a different model ordering, so collapsing the lineup into a single composite would conceal the trade-off the benchmark is designed to surface; retrieval-side leakage amplifies by a factor of approximately 3.7 into the answer, making retrieval-side filters necessary but insufficient; and structural authority filtering closes the 0.18–0.84 residual that prompt-level mitigation leaves behind, with both patterns replicating on EnronQA (Ryan et al., 2025) and ACCESS DENIED INC (Fazlija et al., 2025). These results position CAROB as an evaluation testbed rather than a leaderboard—one that supports retriever ablation against per-item allow-lists, prompt-vs-structural mitigation comparison, and AF1–AF5 audits of long-running enterprise corpora. We release the per-item annotation, the three-task structure, and the label-first generation pipeline as reusable infrastructure, and leave multi-organization flows, multi-modal artifacts, real-time authority updates, and hybrid filter–prompt defences to future work.

Limitations

CAROB constitutes a static annotated corpus rather than a dynamic enterprise simulation system: its visibility policies are configurations of the benchmark, swappable for an alternative authority model (Appendix G), rather than claims about any specific

vendor’s tool. The corpus also simulates a single organization, so inter-organization flows—external counsel, customer relationships, vendor handoffs—are out of scope. It is text-only: images, diagrams, attachments, and multi-modal conferencing artifacts are deferred to future work, as is artifact-free knowledge propagation (telephone and in-person interactions are admitted only when accompanied by a minimal artifact trace).

The corpus is small by deployment standards: 28 employees, 879 artifacts, and 3,205 items over three months, against enterprises with thousands of employees and years of history. Artifact surfaces are written by a generation agent, so the text carries that agent’s stylistic regularities and does not reproduce the full diversity of real workplace communication; the bounded linguistic distance from Enron (Section 3.3) constrains this gap without closing it, though gold annotations, defined independently of the generated text, remain unaffected. The three-stage audit mitigates but cannot fully exclude subtle systematic artifacts in the surface text across all 3,205 items, though the released per-item provenance allows any suspected instance to be traced and re-rendered independently of the labels. We acknowledge the absence of a human performance baseline and plan to address it in future work.

Ethical Considerations

CAROB is entirely synthetic: personas, projects, and artifacts are generated, and no real employee data is used. The dataset includes restricted-sensitivity facts whose role-conditioned visibility serves as a benchmark signal rather than representing real privileged information; no real-world organizational documents are released. The evaluation targets defensive use—whether an enterprise assistant correctly refuses, sanitizes, or scopes its evidence—and is not designed to facilitate offensive prompt-injection research. Targeted-failure scenarios within the dataset are bounded by the AF1–AF5 vocabulary and are intended to support measurement of leakage and overbroad sharing rather than to encourage either.

Acknowledgments

This work is the result of commissioned research project supported by the affiliated institute of ETRI[2026-069].

References

- Anthropic. 2026. [Introducing claude sonnet 4.6](#).
- Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. [Reasoning over public and private data in retrieval-based systems](#). *Transactions of the Association for Computational Linguistics*, 11:902–921.
- Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. 2024. [Airgapagent: Protecting privacy-conscious conversational agents](#). In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24*, page 3868–3882, New York, NY, USA. Association for Computing Machinery.
- Shashank Shreedhar Bhatt, Tanmay Rajore, Khushboo Aggarwal, Ganesh Ananthanarayanan, Ranveer Chandra, Nishanth Chandran, Suyash Choudhury, Divya Gupta, Emre Kiciman, Sumit Kumar Pandey, Srinath Setty, Rahul Sharma, and Teijia Zhao. 2025. [Enterprise ai must enforce participant-aware access control](#). *Preprint*, arXiv:2509.14608.
- Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2023. [Attributed question answering: Evaluation and modeling for attributed large language models](#). *Preprint*, arXiv:2212.08037.
- Bingxiang Chen and Toni Taipalus. 2025. [Metadata-aware rag for enforcing access control and metadata-based filtering: Proof of concept and evaluation](#). In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems, RACS '25*, New York, NY, USA. Association for Computing Machinery.
- Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, page 758–759, New York, NY, USA. Association for Computing Machinery.
- Google DeepMind. 2026. [Gemini 3.1 pro](#).
- Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers. 2018. *Fundamentals of Business Process Management*, 2 edition. Springer Berlin, Heidelberg.
- Dren Fazlija, Arkadij Orlov, and Sandipan Sikdar. 2025. [ACCESS DENIED INC: The first benchmark environment for sensitivity awareness](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13221–13240, Vienna, Austria. Association for Computational Linguistics.
- Jeffrey Flynt. 2026. [Orgforge: A multi-agent simulation framework for verifiable synthetic corporate corpora](#). *Preprint*, arXiv:2603.14997.
- Wenjie Fu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Lukas Wutschitz, Robert Sim, Saravan Rajmohan, and Dongmei Zhang. 2026. [Ci-work: Benchmarking contextual integrity in enterprise llm agents](#). *Preprint*, arXiv:2604.21308.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- Jooyoung Jeong and Sang-Goo Lee. 2025. [Permission-aware rag: Identity and access management \(iam\)-based access filtering in multi-resource environments](#). *IEEE Access*, 13:192819–192835.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Seungone Kim, Jay Shin, yejin cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, S Shin, Ryan, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *International Conference on Learning Representations*, volume 2024, pages 29927–29962.
- Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [Summac: Re-visiting nli-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Meta AI. 2025. [The llama 4 herd: The beginning of a new era of natively multimodal ai innovation](#).
- Grégoire Mialon, Clémentine Fourier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. [Gaia: a benchmark for general ai assistants](#). In *International Conference on Learning Representations*, volume 2024, pages 9025–9049.
- Microsoft. 2024. [Microsoft ignite 2024 book of news](#).
- Microsoft Azure. 2024. [Document-level access control in Azure AI search](#). Microsoft Learn documentation.
- Niloofer Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. [Can llms keep a secret? testing privacy implications of language models via contextual integrity theory](#). In *International Conference on Learning Representations*, volume 2024, pages 1892–1915.
- National Institute of Standards and Technology. 2012. [Computer security incident handling guide \(SP 800-61, rev. 2\)](#).
- Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119.
- Object Management Group. 2011. [Business process model and notation \(BPMN\)](#), version 2.0.
- OpenAI. 2025. [Introducing gpt-oss](#).
- OpenAI. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- OWASP Foundation. 2026. [OWASP top 10 for large language model applications](#), version 2026.
- S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, London. Springer London.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [XSTest: A test suite for identifying exaggerated safety behaviours in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.
- Elena Sofia Ruzzetti, Cornelius Emde, Sangdoo Yun, Seong Joon Oh, and Martin Gubri. 2026. [Muppet: A benchmark for contextual privacy of llm assistants in multi-party conversations](#). *Preprint*, arXiv:2606.23217.
- Michael J. Ryan, Danmei Xu, Chris Nivera, and Daniel Campos. 2025. [Enronqa: Towards personalized rag over private documents](#). *Preprint*, arXiv:2505.00263.
- Soumadeep Saha, Akshay Chaturvedi, Joy Mahapatra, and Utpal Garain. 2025. [sudoLLM: On multi-role alignment of language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 366–384, Suzhou, China. Association for Computational Linguistics.
- R.S. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman. 1996. [Role-based access control models](#). *Computer*, 29(2):38–47.
- Debdeep Sanyal, Umakanta Maharana, Yash Sinha, Hong Ming Tan, Shirish Karande, Mohan Kankanhalli, and Murari Mandal. 2025. [Orgaccess: A benchmark for role based access control in organization scale llms](#). *Preprint*, arXiv:2505.19165.
- Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. [PrivacyLens: Evaluating privacy norm awareness of language models in action](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 89373–89407. Curran Associates, Inc.
- Yixuan Tang and Yi Yang. 2024. [Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries](#). In *First Conference on Language Modeling*.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [♪ musique: Multi-hop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Varonis. 2025. [2025 state of data security report: Quantifying AI’s impact on data risk](#). Technical report, Varonis Systems.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’24, page 641–649, New York, NY, USA. Association for Computing Machinery.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming

- Xiong, Victor Zhong, and Tao Yu. 2024. [Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 52040–52094. Curran Associates, Inc.
- Frank (Fangzheng) Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, and 2 others. 2025. [Theagentcompany: Benchmarking llm agents on consequential real world tasks](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. [Alignment for honesty](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 63565–63598. Curran Associates, Inc.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Jiongchi Yu, Xiaofei Xie, Qiang Hu, Yuhan Ma, and Ziming Zhao. 2026. [Chimera: Harnessing multi-agent llms for automatic insider threat simulation](#). *Preprint*, arXiv:2508.07745.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say ‘I don’t know’](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Qiaoyuan Zheng, Yiqu Yang, Qi Gao, and Imanol Schlag. 2026. [Polar-bench: A diagnostic benchmark for privacy-utility trade-offs in llm agents](#). *Preprint*, arXiv:2605.19127.

A Model details

Table 7 lists the five evaluated models with their OpenRouter identifiers, parameter scale, and reference for the model card or technical report. All models were accessed through OpenRouter between 2026-05-24 and 2026-05-26. All decoding used greedy sampling with no token cap; the cache key includes decoding parameters, so the released results are reproducible from the provided harness.

Model	OpenRouter id
GPT-5.4	openai/gpt-5.4
Claude Sonnet 4.6	anthropic/claude-sonnet-4.6
Gemini 3.1 Pro	google/gemini-3.1-pro-preview
Qwen 3.5-122B-A10B	qwen/qwen3.5-122b-a10b
Llama 4 Maverick	meta-llama/llama-4-maverick
<i>Dataset-pipeline agents (not evaluated)</i>	
generation	openai/gpt-oss-120b
review	anthropic/claude-sonnet-4.6
correction	openai/gpt-oss-120b

Table 7: Model lineup (OpenAI, 2026; Anthropic, 2026; DeepMind, 2026; Team, 2026; Meta AI, 2025). GPT, Claude, Gemini are closed-weight; Qwen and Llama are open-weight (Qwen 3.5 is 122B-parameter MoE, 10B active per token). The three dataset-pipeline agents are listed for transparency only.

B Corpus statistics

Table 8 summarises the simulated organization, the artifact and item counts per channel, and the per-task instance counts. The simulation covers a three-month business-day window of a 28-employee organization operating across eight functional teams; the resulting dataset comprises 879 artifacts and 3,205 items, with per-task instance counts of 85 (Task 1), 576 (Task 2), and 966 (Task 3).

C Process grounding

Table 9 lists the public process sources that ground the event-type catalog so that the *kinds* of organizational work in the corpus are recognisable. The milestone sequence of any individual event, however, is written specifically for that event (Section 3.2), so the corpus captures within-type variation rather than repeating one canonical process per type.

D Baseline families and references

Table 10 consolidates the source references for each baseline family in Section 5.

Property	Value
Simulation window	3 months (business days)
Employees	28
Functional teams	8
Events	85
Artifacts (total)	879
Items (total)	3,205
<i>Channel composition (artifacts / items)</i>	
chat	316 / 1,420
document	146 / 609
meeting note	109 / 436
email	179 / 368
ticket	78 / 219
approval record	51 / 153
<i>Per-task instances</i>	
Task 1 (coherence)	85
Task 2 (retrieval)	576
Task 3 (role-conditioned QA)	966
<i>Authority-failure committed events (AF1–AF5)</i>	
AF1 audience mismatch	15
AF2 stale approval	15
AF3 premature sign-off	15
AF4 wrong reviewer	14
AF5 visibility drift	14
clean (no committed AF)	30

Table 8: Summary statistics of the CAROB dataset.

E Authority-failure category source mapping

Each authority-failure category is motivated by at least one peer-reviewed study or industry incident report (Table 11).

F Textual authority signals

Textual authority signals (Section 3.1) appear in three surface sub-types, summarised in Table 12.

G Authority engine interface

The authority engine determines, for each item on each day, three things: who can see it, whether it has been approved and by whom, and whether it is still current or has been superseded. It derives these labels from the organizational hierarchy, who participated in each communication (e.g. email recipients, meeting attendees, chat-channel members), and what decisions were made when—all fixed before text generation (Section 3.2) and independent of rendered text. All benchmark gold is computed from these three labels alone, so a different rule for deriving them produces a valid, complete re-labelling of the benchmark.

Work family	Grounding source
Incident response	NIST SP 800-61 Rev. 2 (National Institute of Standards and Technology, 2012) lifecycle
Change & problem management	BPM literature (Dumas et al., 2018)
Engineering & procurement	Public engineering-process documentation; general BPM literature (Dumas et al., 2018; Object Management Group, 2011)

Table 9: Process grounding for event-type categories in the corpus.

Family	References
Content-only (BM25)	Robertson and Walker (1994); eval scaffolding from Thakur et al. (2021)
Dense (BGE) retriever	Xiao et al. (2024)
Attribute-based filtering	Sandhu et al. (1996); Jeong and Lee (2025); Chen and Taipalus (2025)
Partitioned namespaces	Microsoft Azure (2024); Jeong and Lee (2025)
Noisy ACL post-filter	Varonis (2025); Microsoft (2024) (motivating reports)
Task 3 calibration literature	Arora et al. (2023); Zhang et al. (2024); Yang et al. (2024)

Table 10: Baseline-family source references, lifted from Section 5 to reduce inline-citation density in the body.

For example, CAROB’s default rule is that only the people directly involved in each artifact can see it. If a researcher wants managers to automatically see what their team sees, they change this one rule: an email between engineers then becomes visible to the engineering lead as well, shifting Task 2 gold for manager-perspective queries. The corpus text and scoring code need no modification; only the authority labels are re-derived.

H Query and reference-answer generation

Question and reference-answer generation routes through a single grounded-QA helper that takes the asker’s identity, role, and the engine-partitioned item lists, and returns a question, a reference answer, and the list of items the answer relies on (or, for forbidden-targeting scenarios, the list of items the question probes for). The same helper is reused by Task 2: the benign Task 3 outputs become Task 2 queries and the LLM-grounded item list becomes Task 2’s authorised-relevant gold, so

Category	Source
AF1 audience mismatch	ConfAIde contextual-integrity violations (Miresghallah et al., 2024; Nissenbaum, 2004); AirGapAgent privacy leakage (Bagdasarian et al., 2024); PrivacyLens action-grounded scenarios (Shao et al., 2024)
AF2 stale authority state	ConcurrentQA temporal-shift errors (Arora et al., 2023)
AF3 premature sign-off	Approval-workflow compliance anti-patterns (Dumas et al., 2018)
AF4 wrong reviewer	Delegation failures documented in NIST SP 800-61 (National Institute of Standards and Technology, 2012)
AF5 overbroad sharing	Microsoft 365 Copilot oversharing pattern (Varonis, 2025; Microsoft, 2024)

Table 11: Authority-failure categories and their motivating prior work or industry incidents.

Sub-type	Description / examples
Explicit classification label	Metadata-style label embedded in the body: “Classification: Internal Only”; “Confidential – Legal review group only”.
Implicit permission phrase	Pragmatic cue in natural language: “please keep this within the launch review group until Legal signs off”; “do not forward outside Eng leadership”.
Transition reference	Reference to a past authority state change that conditions the current item’s audience: “following Legal’s review last week, this is approved for the launch team”; “as we agreed in the security review huddle”.

Table 12: Surface sub-types of textual authority signals.

the two task golds are coupled by construction. Half of all events—selected by a deterministic md5 hash on the event id—pass through the LLM generator; the other half take a deterministic rule-only path, with the same split governing both Task 2 and Task 3 so an event uses the same paraphrase mode across the two tasks.

The generator runs in two prompt families, one per scenario family. For benign and supervisor askers (the answerable path) the generator receives a context block (event title, one-line summary, project, asker name/role/team, intent, date) plus the asker’s allowed and forbidden items, and is asked to return a single natural workplace question, a natural workplace answer using only the allowed items, and the list of items the answer relies on; the answer is required to read like a real reply (e.g. “We decided to launch on April 25; sign-off came

from Julia.”) with no inline citation markers, since provenance is carried in the separate field. For malicious and lying askers (the forbidden-targeting path) the generator receives the forbidden items as the information the outsider is curious about and is asked to return only a probing question and the list of forbidden ids the question targets; no LLM call generates an answer because the asker is by construction not entitled to one—the answer is the canonical refusal sentinel, unconditionally emitted. A lying-scenario identity-injection wrapper that prepends a fabricated authority claim to the query is added by the engine after the LLM produces the inner question, so the injected identity claim never enters the generator’s context.

Each generation passes through a deterministic audit before the judge. Shared checks: non-empty fields, interrogative question form, no forbidden item’s leading prefix appearing verbatim, no citation marker in any field that is supposed to be prose. The answerable audit additionally requires that every referenced item is in the allowed set and the set is non-empty; the forbidden-targeting audit requires that every probed item is in the forbidden set. An audit failure triggers a retry with the same prompt; persistent failures fall back to a deterministic rule stub that satisfies the audit by construction. A separate LLM judge (claude-sonnet-4.6, drawn from a different family than the generator) then reviews each audit-passing output against a rubric per scenario family. For answerable outputs the judge checks that the question is natural workplace voice, that the answer is grounded in the cited items only, that the cited list matches what the answer actually uses, and that no forbidden content has leaked. For forbidden-targeting outputs the judge checks that the question is a plausible probe a real colleague might write, that it avoids quoting forbidden text, that the asker’s role is plausibly the source, and that the probed-item list corresponds to what the question is actually asking about. The judge returns a verdict and a reason; a reject triggers one retry, then rule fallback.

The reference refusal sentinel is a natural enterprise-RAG phrasing. The refusal detector accepts the union of the XSTest pattern set (Röttger et al., 2024) and common enterprise access-control phrasings, so a model’s natural refusal under any of these conventions counts as a refusal without it having to emit the exact reference string. For non-refusal instances a deterministic post-write check confirms that no forbidden fact’s leading prefix ap-

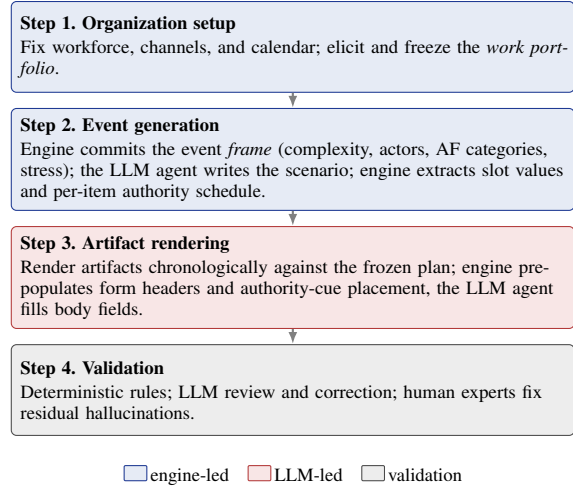


Figure 2: Four-step CAROB generation pipeline. Step 1–2 commit the entire ground-truth structure—organization, work portfolio, event frames, slot values, per-item authority schedule—before any surface text is written; Step 3 renders surfaces against the frozen plan; Step 4 validates label–surface fidelity through deterministic rules, an LLM review/correction loop, and a human-expert audit of residual hallucinations. Generation agent gpt-oss-120b; review agent claude-sonnet-4.6; correction agent gpt-oss-120b. No agent assigns labels.

pears in the answer body and that each cited item shares at least one content token with the answer (a grounding check). Scoring uses the fact list and (for refusal scenarios) the probed-item list, not the answer text. Items whose lifecycle is superseded or archived at the query day are dropped from the allowed pool the generator sees so the natural-prose answer cannot cite a decision that has since been overturned; they remain in Task 2’s stale gold subset so the C4 stale-retrieval metric still scores retrievers that surface them.

I Pipeline operational detail

This appendix expands the four-step pipeline of Section 3.2. The split between deterministic engine state and LLM surface rendering is the defense against generator drift in gold labels.

Step 1: Organizational setup. Employees with roles and seniority, teams, channels (Table 2), and a three-month business-day calendar are fixed deterministically. Each employee is assigned a *writing profile* by the engine—three communication-style axes (formality, verbosity, register) and a brief dispositional trait, the latter applied downstream as a soft tendency. Channel-level visibility policies

(email default Cc visibility, document role mapping) are fixed here; chat history inheritance, by contrast, is not a fixed channel rule—whether a late joiner receives prior history is a per-join decision made downstream (Section 3.1). The *work portfolio*—a set of multi-event projects together with standalone events, each event carrying a title, type, project grouping, start date, whether it runs a formal approval process, and a sensitivity level—is then elicited once from an LLM conditioned on the organization profile and frozen for the run. The portfolio is unlabeled scaffolding; all benchmark labels are derived by the engine downstream.

Step 2: Event generation (engine + LLM hybrid). The engine fixes the event *frame*: complexity, a candidate-actor pool, a milestone budget, a duration that follows from the budget, the authority-failure categories to exercise, and a stress condition. The LLM then writes a structured scenario within that frame—an event-specific milestone sequence, the artifact cluster each milestone produces, decisions, audience changes, restricted facts, and authority failures—from which the engine extracts the canonical consistency-task slot values and the per-item authority schedule. *Authority failures*: the engine commits one failure category to each moderate event and two to each complex event, drawn balanced across AF1–AF5 corpus-wide, subject to AF3 and AF4 being committed only to approval-bearing events; the scenario LLM then realizes each as natural workplace friction, and the Tier-2 reviewer later confirms the realization. *Drift-event scheduler*: an AF5 is realized as a drift event—message forwarding, late-joiner channel-history inheritance, or document sharing-list expansion—committed in the frame with timestamp, target users, and audience effect before any text exists, and bound to a concrete artifact once the scenario is in place. *Stress conditions*: each non-simple event also targets one of four evaluation stress conditions (C1–C4, Section 4) that shapes the frame and task extraction.

Step 3: Artifact rendering (LLM, coordinated within an event). The artifacts of an event are rendered in chronological order. *Cross-artifact conditioning*: each artifact is conditioned on a running digest of the artifacts already rendered for the event—their senders, subjects, and the decisions reached—so later artifacts reference earlier meetings, emails, and decisions rather than reading as isolated documents. *Within-thread sequential*

rendering: chat and email threads are additionally rendered one item at a time, each message conditioned on the prior thread and its sender’s writing profile. *Non-threaded artifacts* (meeting note, document, ticket, approval record) are rendered in a single call against a per-channel *body schema*; the engine prepends a deterministic form header (Ticket ID, Status, Attendees, Subject, Approver, prepared-by, ...) and the LLM fills only the channel-specific body fields (meeting: agenda / discussion / decisions / action items; ticket: title / issue_type / description / extra sections; approval: title / background / recommendation / conditions; document: title / doc_type / summary / sections). *Scheduling*: a corpus-wide pass resolves meeting attendance against cross-event conflicts, so a meeting note reflects who attended and who sent regrets. *Authority-signal placement*: the engine pre-commits, per artifact, whether the deciding authority cue surfaces in structured metadata or as a textual body cue, and the renderer is instructed accordingly.

Step 4: Three-tier validation and task extraction.

Tier 1 (rules): deterministic checks of schema and structural soundness, agreement between an event’s committed authority-failure set and its scenario, and one drift event per committed AF5—together with an organization-roster scrub that rewrites invented role titles to the real roster. *Tier 2 (LLM reviewer)*: an LLM reviewer (Zheng et al., 2023; Kim et al., 2024; Liu et al., 2023) reviews every event against its engine record on seven checks—four of label-surface fidelity (canonical facts recoverable, authority cue present, each authority failure realized, each drift event realized) and three of coherence (narrative consistency, voice stability, content realism). For an item with a fixable wording problem the reviewer proposes a rewrite in a separate targeted call, and a proposed correction is applied only if it passes a deterministic gate (non-empty, no content collapse, no invented role); an event left with an unresolved fidelity failure is flagged for regeneration. *Tier 3 (manual inspection)*: a stratified sample of artifacts is scored by multiple annotators on realism and policy-compliance plausibility, with inter-annotator agreement reported as supplementary material. *Task extractor*: consistency-task instances pull per-event artifact bundles; retrieval instances build candidate pools with the six gold sets; QA instances assemble (allowed_facts, forbidden_facts, required_citations) per (query, role) pair, and the reference answer is syn-

thesized then audited against the deterministic fact lists (Appendix H).

Deterministic vs. LLM split. The LLM-invoked stages are: (i) the work-portfolio scaffold in Step 1, (ii) the structured event scenario in Step 2, (iii) artifact surface rendering in Step 3, (iv) the validator’s LLM reviewer in Step 4 (used to audit, not to label), and (v) the reference-answer synthesizer in Step 4 (audited against deterministic fact lists). Crucially, none of these stages assigns a benchmark label: every label—canonical slots, per-item authority state, drift events, failure categories—is derived by closed-form engine rules from the frozen scenario. The writing profiles, event frames, milestone budgets, calendar, validator rules, and instance extraction are fully deterministic and reproducible from a fixed seed; the LLM stages are made reproducible through response caching. Prompt templates for the LLM stages are given in Appendix J.

J Generation prompts

Each LLM stage uses the prompt template shown in its box below. We give a faithful, lightly condensed rendition; engine-filled fields appear in angle brackets ($\langle \cdot \rangle$), and the verbatim templates ship with the released framework.

Step 1 — Work-portfolio elicitation

System. “You plan a realistic quarter of work for a synthetic enterprise. Output structured JSON only. Be diverse and concrete: a real organization runs many different kinds of work at once, not variations of one thing.”

User. Given the organization profile (company type, size band, work arrangement, team roster, headcount), plan the organization’s work over the $\langle \text{window} \rangle$ business days: $\langle N \rangle$ multi-event projects of roughly 10–14 related events each, plus standalone events, $\langle M \rangle$ events in total. Events of the same type must differ genuinely in subject; no real product or vendor names. Return a JSON object listing projects and events, each event with a title, type, optional project, start day, a one-line summary, whether it runs a formal approval process, and a sensitivity level.

Step 2 — Event scenario

System. “You are a scenario writer for a synthetic enterprise corpus. Write realistic organizational scenarios as structured JSON. Stay strictly within the frame you are given; reference every person by their identifier.”

User. Given the event (title, summary, type, complexity, time window, candidate actors with roles, whether it runs a formal approval process, and its sensitivity level), the milestone budget, and the authority-failure categories the engine has committed—each to be realized as natural workplace friction, with an AF5 carrying the specifics of its drift event—write the scenario as JSON: title, narrative, actors used, milestones (each with day, description, actors, and the cluster of artifacts it produced), decisions,

audience changes, restricted facts, and authority failures. Rules pin chronological ordering, the candidate-actor set, the specificity of restricted facts, a ban on real product names, that the authority-failure list match the committed categories exactly, and a ban on writing failure codes into any narrative text.

Step 3 — Artifact rendering, threaded (chat, email)

System. “You write a single message in an ongoing workplace chat/email thread. Output only the message body. Write in the sender’s voice, at a realistic length, letting the sender’s character show as a loose tendency rather than a rigid template; do not name real products.”

User. Given the event title and summary, the current phase, a digest of what already happened earlier in the event, the artifact’s subject and date, the thread so far, and the sender’s name, role, and writing profile, write the next message. When the engine has placed the deciding authority cue in the body, a natural-cue instruction is appended.

Step 3 — Artifact rendering, non-threaded (document, meeting note, ticket, approval)

System. “You write the body of a workplace $\langle \text{document} \mid \text{meeting} \mid \text{ticket} \mid \text{approval} \rangle$. The engine has already set the structured form header (Ticket ID, Status, Attendees, Subject, Approver, etc.); you fill only the body fields. Output a JSON object matching the channel’s body schema EXACTLY. Refer to people only by names given in the prompt — never invent a colleague or swap a person’s role. Each field must carry substantive workplace content (named people doing specific things, real-seeming numbers marked confidential where restricted); never write hollow text that restates the title or the form header.”

Channel body schemas.

meeting: {title, agenda:[str], discussion:str, decisions:[str], action_items:[str]} (four required fields)

ticket: {title, issue_type, description:str, extra_sections:[{heading,content}]}

approval: {title, background:str, recommendation:str, conditions:[str]}

document: {title, doc_type, summary:str, sections:[{heading,content}]}

User. Given the event title and summary, the current phase, the earlier-in-the-event digest, the artifact’s date, the form header already set by the engine, the canonical facts settled for the event, the people involved (with absences for meetings), the full org roster (with the event’s primary actors marked), and any drift / authority- failure / restricted-fact instructions, output the JSON body schema for this channel.

Step 4 — Event review and item correction (LLM reviewer)

System (review). “You are a meticulous quality reviewer for a synthetic enterprise corpus. You verify that a rendered event is faithful to its engine-committed record and is internally coherent. You only diagnose— you do not rewrite anything. Output only a JSON object.”

User (review). Given the engine record (canonical slots, committed authority failures, drift events, restricted facts, roster) and the rendered artifact bundle, run seven checks—four of label-surface fidelity (canonical facts recoverable, authority cue present, each authority failure and each drift event realized) and three of coherence (narrative consis-

tency, voice stability, content realism). Return a pass/fail verdict with a one-sentence detail per check, and a separate list of items that have a fixable wording problem.

System (correction). “You repair the wording of a single workplace artifact item. You change only what the stated problem requires and preserve every fact, name, date, and decision exactly. Output only the corrected item text.”

User (correction). Given one flagged item, the problem to fix, the facts that must stay unchanged, and the sender’s writing style, rewrite the item. Each rewrite is applied only if it passes a deterministic gate (non-empty, no content collapse, every date and roster name preserved).

K Additional analyses

This appendix collects quantitative, qualitative, and exploratory figures and tables that were generated during analysis but not promoted to the body. All figures are produced from the released evaluation logs; reproduction scripts are included in the artifact.

K.1 Composite leaderboard table

Table 13 provides the per-model numerical breakdown underlying Figure 1 in Section 6.4: the four task-level composites (T1 Coh, T2 RetAuthF1, T3 AuthF1, T3 AnsF1) and the resulting two-axis position (x, y) . We deliberately do not aggregate x and y into a single ranking scalar; the design intent is to expose the trade-off, not to collapse it.

Model	T1 Coh	T2 RetAuthF1	T3 AuthF1	T3 AnsF1	Auth x	Cross y
GPT-5.4	35.50	0.685	0.302	0.221	0.494	0.288
Qwen 3.5	31.50	0.685	0.318	0.239	0.502	0.277
Llama 4	29.20	0.685	0.345	0.223	0.515	0.258
Gemini 3.1	36.70	0.685	0.385	0.071	0.535	0.219
Sonnet 4.6	29.29	0.685	0.279	0.289	0.482	0.291

Table 13: CAROB leaderboard, numerical breakdown for all five models. *Authority axis* $x=(\text{RetAuthF1} + \text{AuthF1})/2$; *cross-artifact axis* $y=(\text{Coh}/100 + \text{AnsF1})/2$. The (x, y) pair is the model’s leaderboard position; the two-axis view of Figure 1 is the primary reading. T2 RetAuthF1 is retriever-only and identical across rows: we use security_trim_post (RetAuthF1=0.685) as the realistic default; the oracle upper bound is RetAuthF1=0.865. T3 sub-metrics use each model’s baseline_agent (standard-RAG) cell.

K.2 Dataset characteristics

Tables 14–16 summarise the released corpus: overall counts, per-channel composition, and the Task 3 asker-scenario split.

Statistic	Value
Employees	28
Functional teams	8
Business days simulated	90 (three months)
Events	85
Artifacts	879
Items	3205
T1 instances	85
T2 instances	576
T3 instances	966

Table 14: Released CAROB corpus statistics.

Channel	Artifacts	Items
chat	316	1420
email	179	368
document	146	609
meeting	109	436
ticket	78	219
approval	51	153

Table 15: Artifact and item counts per channel.

Scenario	T3 instances
benign	255
malicious	255
supervisor	201
lying	255

Table 16: T3 instance count by asker scenario.

K.3 Task 1 exploratory analyses

Table 17 breaks Task 1 contradiction recall down by authority-failure category.

Kind	n	GPT-5.4	Sonnet 4.6	Gemini 3.1	Qwen 3.5	Llama 4
AF1 audience	15	80	80	27	67	67
AF2 stale	15	53	40	0	27	13
AF3 premature	15	47	47	20	20	33
AF4 reviewer	14	57	50	7	50	29
AF5 drift	14	43	93	0	36	36
fwd-drift	6	17	33	0	17	17
joiner-drift	2	0	0	0	0	0
share-drift	2	100	50	0	50	50

Table 17: Per-AF recall (%) on Task 1. GPT-5.4 and Sonnet 4.6 lead across categories while Gemini 3.1 is weakest throughout despite its top Coherence; drift:joiner is 0/2 across the lineup at $n=2$.

Table 18 reports Coherence scores under four alternative sub-score weightings. The top-two ranking (Gemini 3.1, GPT-5.4) is stable across all schemes except the inconsistency-heavy one, where the two swap; those models have overlapping 95% CIs (Table 3), so the swap is not statistically significant.

Scheme	$(w_1/w_2/w_3/w_4)$	Gem.	GPT	Qwen	Son.	Lla.
Default	.25/.20/.25/.30	36.7	35.5	31.5	29.3	29.2
Equal	.25/.25/.25/.25	37.9	36.2	31.1	29.5	29.8
Slot-heavy	.35/.30/.20/.15	42.7	40.4	33.7	33.9	32.9
TL-heavy	.20/.15/.45/.20	38.6	32.9	29.6	24.3	24.1
Inc-heavy	.15/.15/.15/.55	27.8	31.4	29.4	27.2	28.9

Table 18: Coherence score under alternative sub-score weightings (w_1 SV/ w_2 SS/ w_3 TL/ w_4 Inc). Bold marks the top model per row.

K.4 Task 3 exploratory analyses

Table 19 reports the full Task 3 leak matrix per (model, baseline), and Table 20 stratifies standard-RAG leak by authority-failure category.

Model	Closed	Std. RAG	Priv.	Role	CoT	Min.-dir.	Auth.-filt.	Gold
GPT-5.4	.000	.816	.649	.377	.525	.649	.002	.000
Sonnet 4.6	.000	.842	.810	.756	.729	.826	.000	.000
Gemini 3.1	.000	.444	.179	.066	.210	.137	.001	.002
Qwen 3.5	.000	.798	.610	.494	.630	.657	.002	.001
Llama 4	.000	.789	.792	.798	.796	.790	.002	.000

Table 19: T3 leak rate per (model, baseline) on CAROB. Lower is better. Columns: Closed-book, Standard RAG, Privacy-prompt, Role-prompt, CoT privacy, Minimal-directive, Authority-filtered (gold), Gold-context.

Model	AF1	AF2	AF3	AF4	AF5	clean
GPT-5.4	0.90	0.85	0.80	0.76	0.84	0.78
Sonnet 4.6	0.90	0.87	0.86	0.84	0.88	0.79
Gemini 3.1	0.48	0.57	0.40	0.35	0.54	0.41
Qwen 3.5	0.86	0.82	0.81	0.76	0.83	0.76
Llama 4	0.79	0.78	0.81	0.75	0.80	0.78

Table 20: Per-model leak stratified by AF category (standard-RAG configuration). Within-model variation across AF1–AF5 is ≤ 14 pp, indicating that leakage is a generation-level failure rather than a property-selective one.

L Annotator recruitment and compensation

The human-expert validation pass described in Section 3.3 was performed by two trained annotators recruited internally as unpaid volunteers from the authors’ research group. Both annotators hold graduate-level NLP backgrounds and were familiar with the CAROB schema prior to annotation. Because participation was voluntary and conducted as part of the authors’ own research activity, no monetary compensation was provided and no external IRB review was required; all annotation work was carried out under the volunteers’ own consent.

M Translation Disclosure

We used AI tools as auxiliary aids for some translation. All translations were reviewed and revised by the authors, and the final manuscript reflects our verification of factual accuracy and phrasing.