IR-AGENT: DECOUPLED ALIGNMENT FOR LARGE LANGUAGE MODEL DRIVEN EARNINGS CALL RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Company-side earnings-call answering demands three concurrent desiderata: every cited figure must be grounded in the company's publicly disclosed materials, relevant available figures should be properly cited in replies, and responses should adopt managerial conversational register instead of plain document recitation. Satisfying all three simultaneously poses a substantial challenge for large language models (LLMs), as these objectives are inherently conflicting: models optimized to mimic human-like speech tend to invent unsupported numeric claims. In this paper, we first propose IR-Agent, an industry agent for high-fidelity, disclosure-consistent earnings-call response generation, whose architecture is the Decoupled Alignment Framework (DAF). The style module aligns the agent to authentic managerial utterances via supervised fine-tuning, with style signals deliberately excluded from reward computation to avoid reward hacking. The disclosure module regulates numeric citation behavior through advantage-weighted policy optimization, exclusively tuning numerical grounding preference while preserving learned linguistic style. The verification module leverages a rule-based numeric verifier to filter and rank sampled candidates, eliminating fabricated numbers; this same verifier also provides consistent reward signals during training. We construct IRCorpus, the first company-side earnings-call response corpus with verifier-supported numeric supervision, and devise a closed-form evaluation protocol that quantifies fabrication, omission and stylistic fidelity without relying on LLM-as-judge metrics. On 573 held-out questions from 20 companies unseen during fine-tuning, eight-candidate IR-Agent, built on a 4B-parameter base model, yields zero verifier-detected fabricated numbers with high usability of grounded figures, and matches human managerial register more closely than eight frontier baselines, which persistently suffer from fabrication or excessive omission.

1 INTRODUCTION

Earnings-call question answering is a primary investor-relations channel for public companies: management responds to analyst inquiries about financial performance and business outlook, on the record (Frankel et al., 1999). Unlike general-domain dialogue, this scenario is governed by regulatory and pragmatic constraints. Under Regulation FD and comparable disclosure rules, a company may not release previously unpublicized material information, so the numbers appearing in an answer are expected to trace back to what the company has already disclosed. Beyond that constraint, the financial and accounting literature shows that managerial language acts as a market signal: mechanically excerpted, rigid or inconsistent responses are associated with lower perceived transparency and greater information uncertainty (Bushee et al., 2018; Gow et al., 2021; Hollander et al., 2010; Cazier et al., 2021). Bushee et al. (2018) separate complexity that transmits information from complexity that obscures it, so the same surface form can read as either; Gow et al. (2021) and Hollander et al. (2010) show that non-answers and silence on calls are measurable and consequential disclosure choices rather than conversational noise; and Cazier et al. (2021) find that lengthy boilerplate risk language is assessed unevenly by courts and regulators.

This scenario imposes three desiderata that pull against one another: factual rigor demands that no number be unsupported by the disclosed materials; pragmatic communication requires that valid disclosed figures actually be cited rather than avoided; and relational consistency enforces a human managerial conversational register rather than document-style recitation. The core challenge lies

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in the tension among them. Models optimized to imitate fluent human speech tend to produce plausible but fabricated numbers, the failure that truthfulness and factual-precision benchmarks were built to expose (Lin et al., 2022; Min et al., 2023), while fact-centric paradigms typically produce stiff, excerpt-heavy replies that violate real-world investor-relations communication norms. Existing financial NLP and alignment methods do not resolve this conflict. Conventional joint optimization strategies that blend stylistic and factual objectives within a unified loss or reward are exposed to reward hacking, where the update exploits superficial statistical patterns instead of learning the underlying communicative property (Amodei et al., 2016; Skalse et al., 2022; Gao et al., 2023a). Numerical support, numerical participation and register are different properties of an answer, and we treat them as different.

The mechanisms we build on are established ones: imitation followed by preference optimization (Ouyang et al., 2022; Rafailov et al., 2023), advantage-based policy updates over sampled groups (Shao et al., 2024; Liu et al., 2025; Ahmadian et al., 2024), and generating several candidates before selecting among them with a verifier (Cobbe et al., 2021; Lightman et al., 2024). We propose IR-Agent, a three-module Decoupled Alignment Framework (DAF) that separates style distribution learning, numerical behavioral regulation and factual verification into independent stages. First, the style module applies supervised fine-tuning to capture authentic managerial linguistic characteristics, with no style term entering any later reward and with the demonstrations filtered to answers whose numbers are supported by the materials. Second, the disclosure module performs advantage-weighted policy optimization under a reward that contains only supported and unsupported number counts, tuning numerical citation preferences without perturbing the learned linguistic style; our main configuration uses RLOO. Third, the verification module applies a rule-based numeric verifier to filter and rank multi-sample candidates, removing numerically unsupported answers while supplying the same counts as consistent training rewards. The verifier tests numeric membership, not legal compliance.

To avoid subjective evaluation bias, we construct a closed-form evaluation protocol that involves no model as a judge and quantifies fabrication rate, omission rate and stylistic fidelity, in the spirit of attribution and factual-precision measurement (Rashkin et al., 2023; Gao et al., 2023b), and we apply it to IRCorpus, our company-disjoint corpus of company-side answers.

We conduct experiments against eight strong baseline models, each given one response per question. Eight-candidate IR-Agent achieves zero verifier-detected fabricated numbers with high usability of grounded figures, and aligns more closely with real managerial register than the baselines on the surface proxies we measure; it does not lead on every metric, since five comparators omit less than we do and supervised fine-tuning copies less. Overall, our decoupled paradigm provides a practical and robust solution for high-stakes, style-sensitive and disclosure-constrained financial generative tasks.

Contributions. We summarize our core contributions as follows.

- **A decoupled optimization paradigm.** We identify the conflict between stylistic alignment and numerical grounding in company-side answer generation, and adopt a decoupled principle that assigns style distribution learning and numerical behavioral regulation to separate mechanisms rather than mixing them into one loss or reward, which is where joint optimization becomes exposed to reward hacking (Amodei et al., 2016; Skalse et al., 2022).
- **A three-module IR-Agent.** IR-Agent realizes the principle as the Decoupled Alignment Framework (DAF): supervised fine-tuning learns managerial style from supported demonstrations with no style term in any later reward, policy optimization is confined to numerical grounding behavior, and a rule-based verifier filters and ranks candidates, removing verifier-detected fabricated numbers from the delivered answer while supplying the same counts as training rewards. We pair the system with a closed-form, judge-free protocol that measures fabrication, omission and stylistic fidelity.
- **A corpus and two empirical findings.** We construct a company-split corpus whose supervision is restricted to answers supported by the disclosed materials. Ablations yield two results that run against expectation: the operating point is set by the reward rather than by the choice of advantage estimator (under one reward, GRPO, Dr. GRPO and RLOO give 0.000–0.003 unsupported numbers and 11.9%–15.7% omission, while changing the reward

and its schedule reaches 3.1% omission); and system-level fabrication depends on whether the pool contains a clean candidate rather than on average single-sample quality.

Section 2 positions the work, Section 3 describes DAF, Section 4 presents the experiments, and we close with limitations and conclusions.

2 Related Work

2.1 Financial NLP and Earnings-Call Analysis

Work on earnings calls is predominantly discriminative. Tone and readability are measured with finance-specific dictionaries (Loughran & McDonald, 2011); silence and non-answers are analyzed as disclosure choices to be predicted rather than as conversational noise (Hollander et al., 2010; Gow et al., 2021); EvasionBench supplies a three-level taxonomy of managerial evasion over a large collection of question–answer pairs (Ma et al., 2026), and recent work extends the construct to textual and vocal cues (Kim et al., 2026; Chen & Zhang, 2026). SubjECTive-QA annotates clarity, relevance, specificity and caution in company answers that already exist (Pardawala et al., 2024). Each of these resources labels or scores disclosures that have already been made.

Generative work on the same material addresses a different speaker or a much shorter output. ECTSum pairs transcripts with expert bullet-point summaries (Mukherjee et al., 2022); recent systems draft analyst-side analytical reports (Goldsack et al., 2024) or generate analyst questions (Juan et al., 2024); and FinQA and TAT-QA pose numerical reasoning over financial evidence, returning short extracted or computed answers (Chen et al., 2021; Zhu et al., 2021). To our knowledge, no prior work studies generation of the company-side answer under simultaneous constraints on numerical grounding and managerial register. Producing that answer, and evaluating it in a way that separates candidate cleanliness, candidate-pool coverage and omission, is what this paper contributes.

2.2 LLM Alignment and Multi-Objective Optimization

Standard alignment couples imitation with preference optimization: supervised fine-tuning installs a target style, and a reward then shapes behavior (Ouyang et al., 2022; Rafailov et al., 2023). When several qualities are folded into one reward, optimization pressure concentrates on whichever of them is cheapest to measure. This failure is documented as reward hacking (Amodei et al., 2016), characterized in terms of proxy divergence (Skalse et al., 2022), and measured for reward models as overoptimization (Gao et al., 2023a). Surface style is especially exposed to it, because style statistics are cheap to score and can be satisfied without the communicative property they stand for. We instead assign the three requirements to three mechanisms, and hold each mechanism to the quantity it can actually measure.

Within the alignment family we compare advantage constructions rather than proposing a new one: group-relative advantages (Shao et al., 2024), a bias-corrected variant (Liu et al., 2025), and leave-one-out baselines (Ahmadian et al., 2024), all inside a single offline likelihood update with the rollout pool, starting checkpoint and schedule held fixed. In our measurements the estimator barely moves the result, whereas the shape of the reward moves the operating point substantially. The reasoning for assigning responsibilities to separate mechanisms follows multitask principal–agent analysis, in which rewarding a measurable dimension diverts effort from the less measurable ones (Holmström & Milgrom, 1991). We use that as design motivation rather than as derivation: decidable, measurable and imitable properties overlap instead of forming a partition.

Selecting among several candidates with a verifier is a distinct line of work. Self-consistency and best-of-n ranking aggregate or score samples (Wang et al., 2023; Lightman et al., 2024; Cobbe et al., 2021); attribution research measures whether generated statements are supported by their sources (Rashkin et al., 2023; Gao et al., 2023b); and factual-precision benchmarks score long-form output at fine granularity (Lin et al., 2022; Min et al., 2023). Our verifier is a deterministic numeric-membership test rather than a learned scorer, which yields an exact decision for a restricted property and exposes failure when no candidate passes. System-level reliability then depends on whether the pool contains a clean candidate rather than on average candidate quality. Because the same rule also supplies the training reward, the guarantee is bounded by the rule and by candidate coverage, and independent evaluation matters more rather than less.

3 METHOD

IR-Agent assigns management register, numerical disclosure behavior, and output numerical consistency to three modules: demonstration learning (L1), reward shaping (L2), and inference-time verification (L3), matching the labels in Figure 1. The figure shows the policy progressing from the base model π_0 to π_{SFT} and then π_{IR}, which is frozen during final response selection.

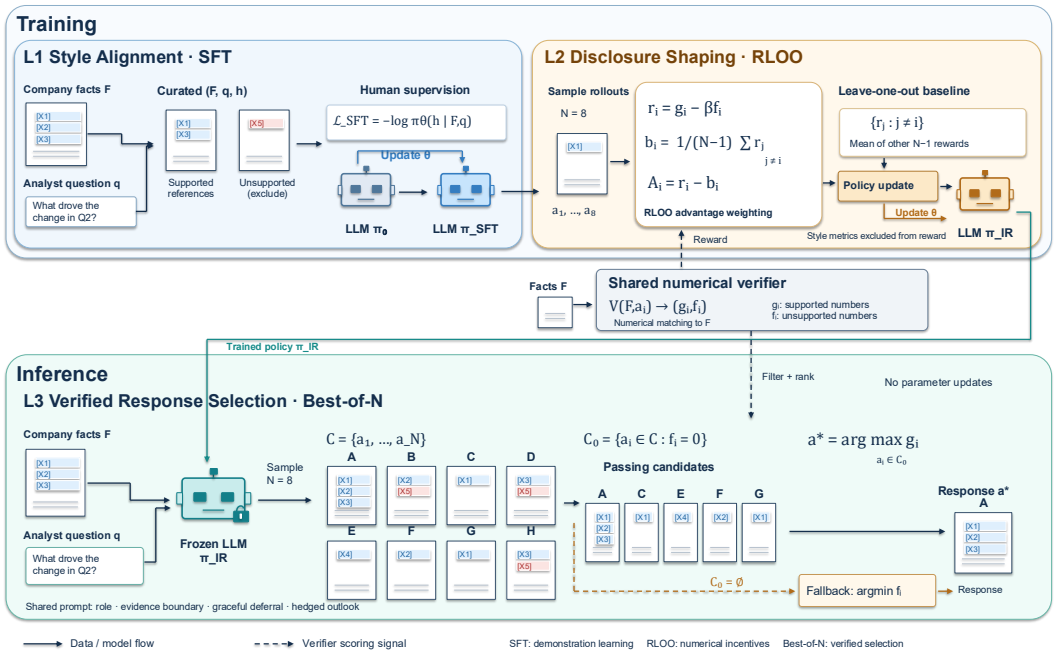

Figure 1: **IR-Agent and the Decoupled Alignment Framework.** L1 learns from curated materials F, questions q, and human answers h; L2 scores a fixed rollout pool with the shared verifier and updates the policy with RLOO advantages; L3 samples, filters and ranks responses without parameter updates, falling back to a minimum-violation candidate if none passes. One rule-based verifier serves L2 and L3, and checks numeric-string membership rather than semantic entailment.

Task and shared verifier. An input is (F, q), where F contains structured financial facts and retrieved prepared remarks. A human answer h is available for supervision and reference-based evaluation but is not supplied to the generator at inference. Let $\mathcal{N}(s)$ be the set of extracted numeric strings in text s. The verifier returns supported and unsupported counts:

$$V(F, a) = (g, f), \qquad g = |\mathcal{N}(a) \cap \mathcal{N}(F)|, \quad f = |\mathcal{N}(a) \setminus \mathcal{N}(F)|. \tag{1}$$

The rule removes commas and leading zeros, retains strings with at least two remaining digits, and counts distinct strings. It does not canonicalize units or associate a value with its financial item or period. Thus, "fabrication" below means an *operational unsupported-number count*, not a complete truth judgement. During training, (g, f) determines the reward; at inference it determines the filter and ranking key. Style metrics are evaluation diagnostics, not reward terms or selection criteria. Our policies are queried with a fixed system instruction that states the role and the evidence boundary and explicitly forbids inventing numbers that are not in the supplied facts (Appendix A.4); detected unsupported numbers are therefore measured with that instruction in place, not as spontaneous behavior. The record does not document whether the API comparators received the identical instruction, so instruction parity across systems is an open caveat rather than an established control.

3.1 L1: STYLE ALIGNMENT

The first module learns how management responds. Demonstrations satisfy both $g(F, h) \geq 1$ and $f(F, h) = 0$. The second condition does the work: admitting one supported number alone would still

allow unsupported numbers elsewhere in the same answer. Filtering reduces the mismatch between visible materials and target answers without assuming that the trained policy inherits the targets' numerical consistency.

We adapt the base model with LoRA and answer-token cross-entropy, masking prompt and padding positions. The experiment record identifies the starting policy as v4/checkpoint-15 and records eightfold numeric-token weighting; we report that weighting as recorded and do not isolate its contribution in our experiments. The implemented supervised objective therefore has the form

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(F,q,h)}\left[\frac{1}{Z_h}\sum_t w_t \log \pi_\theta(h_t \mid F, q, h_{<t})\right], \tag{2}$$

where w_t is the token weight and Z_h the implementation's normalization. Figure 1 uses likelihood shorthand; Appendix A identifies the missing launch-manifest details needed for exact checkpoint reproduction.

Style stays out of the reward because a suffix count or deferral pattern does not fully specify clear, responsive communication. This is a design rationale, not a tested claim that style rewards necessarily fail. Supervised imitation can also change numerical behavior, so we explicitly measure it rather than interpreting this module as affecting style alone.

3.2 L2: Disclosure Shaping

We sample $K = 8$ responses per prompt from π_{SFT} for 1,000 training prompts, producing a fixed pool of 8,000 responses. The main configuration uses

$$r_i = g_i - \beta f_i, \quad \beta = 2; \qquad b_i = \frac{1}{K-1}\sum_{j\neq i} r_j; \qquad A_i = r_i - b_i. \tag{3}$$

We then minimize signed advantage-weighted answer-token negative log-likelihood:

$$\mathrm{NLL}_i(\theta) = -\frac{1}{T_i}\sum_t \log \pi_\theta(a_{i,t} \mid F, q, a_{i,<t}), \qquad \mathcal{L}_{\mathrm{DAF}}(\theta) = \frac{1}{M}\sum_i A_i \,\mathrm{NLL}_i(\theta). \tag{4}$$

The sign increases the likelihood of positive-advantage answers and decreases that of negative ones. Responses and advantages remain fixed during updates; there are no refreshed on-policy samples, clipped likelihood ratios, or learned value model. This is offline policy adaptation, not a new online RL algorithm or a guaranteed policy-improvement procedure.

We compare group-standardized advantages $(r_i - \bar{r})/(s_r + 10^{-4})$ and centered advantages $r_i - \bar{r}$, labeled GRPO-style and Dr.-GRPO-style, respectively. With fixed K, $A_i^{\mathrm{RLOO}} = K(r_i - \bar{r})/(K-1)$; similarity to group centering is therefore partly structural. An anchored reward $r_{\mathrm{anchor}} = -|g - g_h| - \beta f$ instead targets the supported count g_h of the human training answer. Only this anchored variant needs the reference during reward calculation. A precision-weighted reward and the RFT control are specified in Appendix A.7; neither the main reward nor inference selection uses h.

3.3 L3: Verified Response Selection

The frozen policy generates $\mathcal{C} = \{a_1, \ldots, a_N\}$; at evaluation $N = 8$, with temperature 0.8, top-p 0.95 and a 300-token limit. The verifier constructs $\mathcal{C}_0 = \{a \in \mathcal{C} : f(F, a) = 0\}$. We deliver

$$a^* = \begin{cases} \arg\max_{a\in\mathcal{C}_0} g(F,a), & \mathcal{C}_0 \neq \varnothing, \\ \arg\min_{a\in\mathcal{C}} f(F,a), & \mathcal{C}_0 = \varnothing. \end{cases} \tag{5}$$

Ties follow candidate order. The fallback returns a response even when all candidates violate the rule; it is not abstention. We report this branch's frequency as the *all-dirty-pool count* and its complement as the *usable rate*. A response with no extracted number can pass, so cleanliness does not establish useful participation.

Conditional property. For a nonempty pool checked with the selection verifier, $f(F, a^*) = 0$ if and only if $\mathcal{C}_0 \neq \varnothing$. This follows directly from Equation 5 and implies

$$\mathbb{E}[f(F, a^*)] = \Pr(\mathcal{C}_0 = \varnothing)\,\mathbb{E}\left[\min_{a\in\mathcal{C}} f(F,a) \mid \mathcal{C}_0 = \varnothing\right]. \tag{6}$$

Table 1: Final company-disjoint splits. Every reference in these selected subsets has at least one extracted number, with all extracted numbers supported by the materials.

	Train	Validation	Test
Instances	2,252	260	575
Companies	186	13	20

The property requires no independence assumption between candidates. A policy can be clean on average while repeatedly failing on particular questions; selection cannot repair an all-dirty pool. Pool availability must therefore be measured directly.

A simple incentive model. Suppose stating a number gains one unit if supported and loses β otherwise, and the policy assigns subjective support probability p. Its expected incremental reward is $p - \beta(1-p) > 0$ only when $p > \beta/(1+\beta)$. Increasing the penalty raises the threshold for participation. This illustrates how fewer errors can accompany more omission; it is not a theorem about training or a claim that training cannot improve factuality.

4 Experiments

4.1 Dataset

We pair analyst questions with management answers and company materials assembled from structured financial facts and prepared remarks. Retrieval uses the company, reporting period, question, and financial-item terms to retain up to 40 eligible sentences, restored to source order. The human answer is used for filtering and supervision, not as the retrieval query. Materials are stored in each prompt; generation does not execute live retrieval.

Companies are split before example selection (Table 1). The final 2,252 training examples are the numeric-answer subset of 3,895 length-filtered examples. Both validation and test are also filtered to require $g(F,h) \geq 1$ and $f(F,h) = 0$, yielding a benchmark for supported numerical participation rather than a representative sample of all earnings-call questions. Every retained test reference passes the numerical check by construction. Company-disjoint fine-tuning does not establish absence from base-model pretraining.

The full test set contains 150 FORECAST, 132 CAUSE, 114 NUM, 83 SCOPE, 41 COMPARE, 31 EVIDENCE, and 24 OTHER questions (26.1%, 23.0%, 19.8%, 14.4%, 7.1%, 5.4%, and 4.2%). These are construction labels, not newly collected human annotations. Development used validation data, but test ablations were also inspected; we therefore describe company-disjoint evaluation without asserting a preregistered, untouched final test.

4.2 Experimental Settings

Training and inference. We adapt Qwen3-4B-Instruct-2507, a 4B-parameter base model, with LoRA, starting disclosure shaping from the recorded SFT step-15 checkpoint. The fixed rollout pool contains 1,000 training prompts with eight answers each. The estimator comparison uses AdamW, learning rate 10^{-5}, batch size eight, gradient clipping at 1.0, and 60 updates. Local generation uses temperature 0.8, top-p 0.95, and at most 300 new tokens. The main system selects from eight candidates; the N sweep uses nested prefixes of the stored generations. Appendix A distinguishes recoverable settings from missing SFT launch metadata.

Comparators and coverage. We evaluate eight API-served models alongside the base, SFT, and RLOO policies. Six API models and all local policies cover 575/575 questions. Grok and GLM cover 574/575, missing zero-based indices 457 and 101, respectively; their intersection contains 573 questions. All test comparisons use this common set. Responses recorded as errors or empty outputs are excluded, as are answers under 15 words whose reasoning-token count nearly exhausts the output budget; substantive refusals are kept and missing responses are never imputed. Model identifiers

Table 2: Common test set (573 questions); all thirteen metrics are closed-form, recomputed from saved text with no LLM judge, and specified in Appendix A.3. N is the number of candidates available to selection: Fab., Grd., Prec., Miss, Use. and Cln. are grounding and availability measures; Nom., Defl., Copy, Mech. and Act. are surface proxies; Emp. and Rep. are degenerate-output checks. Human references satisfy $f = 0$ by construction, so precision is 1.000 and there is no candidate pool; the grounded count and the style proxies are read against the human reference rather than minimized uniformly.

Method	N	Fab.↓	Grd.	Prec.↑	Miss↓	Use.↑	Cln.↑	Nom.	Defl.	Copy	Mech.	Act.	Emp.↓	Rep.↓
Human	—	0.000	1.417	1.000	—	—	—	1.970	0.011	0.461	0.140	0.923	0.00	0.427
Claude-Opus-5	1	0.138	3.159	0.958	10.6	89.2	89.2	3.653	0.079	8.256	0.150	0.981	0.00	0.326
GPT-6-Astra	1	0.047	1.251	0.964	46.9	95.5	95.5	4.907	0.088	5.081	0.007	0.716	0.00	0.263
Gemini-3.1-Pro	1	0.080	3.344	0.977	8.4	92.7	92.7	3.353	0.206	19.297	0.014	0.913	0.00	0.332
DeepSeek-Flash	1	0.117	4.049	0.972	13.8	90.2	90.2	3.223	0.092	18.574	0.019	0.668	0.00	0.351
Qwen3.8-Max	1	0.106	2.705	0.962	34.2	91.6	91.6	4.198	0.116	9.970	0.007	0.756	0.00	0.314
Grok-4.7	1	0.108	2.436	0.957	31.1	91.6	91.6	3.216	0.172	11.275	0.009	0.639	0.00	0.292
Kimi-K3	1	0.113	4.515	0.975	8.9	90.8	90.8	3.546	0.110	13.806	0.112	0.885	0.00	0.381
GLM-5.3	1	0.105	5.094	0.980	6.6	91.3	91.3	3.139	0.141	17.354	0.101	0.908	0.00	0.384
Base alone	1	0.127	2.086	0.942	27.9	90.1	90.1	4.745	0.129	7.960	0.021	0.785	0.17	0.335
SFT alone	1	0.353	0.766	0.685	57.1	75.7	75.7	1.656	0.020	1.446	0.161	0.935	0.70	0.456
SFT+RLOO alone	1	0.206	0.747	0.784	59.3	85.5	85.5	1.742	0.018	2.280	0.150	0.925	0.35	0.420
Base + selection	8	0.054	3.161	0.983	14.7	95.3	90.0	4.636	0.115	8.756	0.028	0.812	0.00	0.343
SFT + selection	8	0.000	1.750	1.000	17.5	100.0	78.1	1.753	0.019	2.423	0.183	0.948	0.52	0.464
IR-Agent (full)	8	0.000	1.846	1.000	15.2	100.0	83.4	1.697	0.011	2.656	0.147	0.930	0.17	0.442

follow the experiment record. The API comparators contribute one answer each while the selected local systems use eight: this is a configuration comparison, not an equal-compute evaluation.

Evaluation protocol. Fabrication is mean f; precision is the micro-aggregated $\sum g/(\sum g + \sum f)$; the grounded count is mean g. Omission is the fraction with $g(F, a^*) = 0$ among questions where $g(F, h) \geq 1$; all 573 common questions qualify. Usable rate is the proportion of pools containing any $f = 0$ candidate, whereas candidate cleanliness counts clean responses before selection. The style proxies and the degenerate-output checks, with their implementations, are specified in Appendix A.3. Table 2 reports all thirteen quantities, and Figure 2 shows the two headline quantities. Appendix A.3 specifies these proxies and additional diagnostics. Every score is recomputed from saved text without an LLM judge. Deterministic scoring does not remove sampling variance or metric limitations; we report descriptive estimates, not statistical significance.

4.3 Main Results

Single-response reliability is not the system-level result. GPT-6-Astra produces a clean answer on 95.5% of common questions, but pool-average cleanliness across IR-Agent's first eight candidates is 83.4%; neither equals the first-candidate usable rate of Table 2. SFT alone has 57.1% omission and 0.353 unsupported numbers per first response; its much lower 17.5% omission occurs *only after* eight-candidate selection. RLOO alone remains highly withholding (59.3% omission). These results rule out attributing the final system's reliability to the trained policy alone.

Verified output and pool availability. Eight-candidate IR-Agent delivers zero detected unsupported numbers with 100% usable rate, while API single responses produce 0.047–0.138 unsupported numbers per answer. Base plus selection still has 27 all-dirty pools and 0.054 unsupported numbers per selected answer. SFT plus selection also reaches zero and 100% on the common test set, so these two results are not unique to the full framework. On validation, however, SFT plus selection has one all-dirty pool out of 260, whereas RLOO plus selection has none (Appendix A.5).

Quantitative participation. IR-Agent omits on 87/573 questions (15.2%), compared with 100/573 (17.5%) for SFT plus selection; the mean grounded count rises from 1.750 to 1.846. This is a 2.3-point test-set improvement, not a uniform gain: on validation, omission is 20.8% for RLOO plus

Table 3: Nested candidate-count ablation on the same 573 questions. Parentheses give the raw number of all-dirty pools; omission (Miss) is reported separately.

Policy	Measure	$N = 1$	$N = 2$	$N = 4$	$N = 8$
Base	Fab. (dirty)	0.127 (57)	0.082 (39)	0.070 (34)	0.054 (27)
	Miss (%)	27.9	22.3	18.2	14.7
SFT	Fab. (dirty)	0.353 (139)	0.087 (45)	0.005 (3)	0.000 (0)
	Miss (%)	57.1	49.7	33.0	17.5
SFT+RLOO	Fab. (dirty)	0.206 (83)	0.045 (20)	0.002 (1)	0.000 (0)
	Miss (%)	59.3	45.9	32.6	15.2

selection versus 16.9% for SFT plus selection. GLM and Gemini also have lower test omission (6.6% and 8.4%) than IR-Agent, while GPT-6-Astra and Qwen omit on 46.9% and 34.2%. The full system combines a clean selected output with intermediate participation; it is not best on omission alone.

Surface register. IR-Agent's nominalization, deflection, and copying scores are 1.697, 0.011, and 2.656%, against human values of 1.970, 0.011, and 0.461%. API models span 3.139–4.907, 0.079–0.206, and 5.081–19.297%, respectively. The full system is closer to the human reference on each proxy than any API model. SFT plus selection is also competitive and is slightly closer than IR-Agent on nominalization and copying. These are surface-profile results, not human preference judgements. GPT illustrates low quantitative participation, whereas Gemini combines low omission with high source overlap. They exemplify different trade-offs, without establishing intentional strategies or an exhaustive two-class taxonomy of models.

Participation, length and degenerate output. The same table separates the remaining quantities. The full system cites more disclosed numbers than the human reference (1.846 against 1.417), and its mechanism-explanation rate is closer to the human value than any API model's (0.147 against 0.140, where the closest API model reaches 0.150). Precision is at ceiling for SFT plus selection and for the full system, and no API model exceeds 0.980. Length is where the system departs most from human behavior: 119.8 words against 220.4, so the delivered answer is about half as long as the reference. Degenerate output is rare but not absent: 0.17% of delivered answers fall under twenty characters against 0.00% for the human reference, and the repeated-token fraction is 0.442 against 0.427. Single responses remain a different picture: SFT alone reaches 0.685 precision and SFT+RLOO alone 0.784, far below what the same policies reach after selection.

4.4 Ablation Study

Candidate count. Table 3 reports nested prefixes. For IR-Agent, increasing N from one to two reduces all-dirty pools from 83 to 20; at four, one remains; at eight, none remains. Omission falls from 59.3% to 15.2%. Eight is the smallest *tested* size yielding zero detected unsupported numbers for this policy on this set, not a guarantee for future questions.

Pool-average cleanliness across the base policy's first eight candidates is 90.0%, above IR-Agent's 83.4%, yet 27 base pools remain entirely dirty at $N = 8$. Aggregate candidate cleanliness therefore does not determine per-question availability. The observation is consistent with question-dependent failure patterns; it does not by itself identify semantic diversity as their cause.

Reward shape and estimator. Under the same linear reward and 60-step schedule, the three estimators give unsupported counts of 0.000–0.003 and omission of 11.9–15.7% (Table 4). RLOO and group centering are especially similar, consistent with their constant-factor relationship. Under the matched GRPO-style schedule, an anchored $\beta = 1.5$ reward changes omission from 11.9% to 11.2%, and unsupported counts from 0.003 to 0.010. The longer anchored $\beta = 1.0$ run reaches 3.1% omission with 0.047 unsupported numbers, but changes duration as well as reward. It establishes another operating point, not the isolated effect of reward shape. Development RFT controls also achieve zero after selection, showing that this property does not require RLOO (Appendix A.7).

High-penalty degeneration. An archived exploratory run combining $\beta = 100$ with token-level numerical credit (`hardtok-120`) reports 0.085 supported numbers, 0.023 unsupported numbers,

Table 4: Estimator and reward configurations, all with $N = 8$ selection on the common test set. The first three rows share the linear reward and 60-step schedule. The last row also changes training duration, and is not a controlled reward-only ablation.

Configuration	Fab.↓	Miss (%)↓	Nom.
GRPO-style, linear, 60	0.003	11.9	1.672
Dr.-GRPO-style, linear, 60	0.000	15.7	1.777
RLOO, linear, 60	0.000	15.2	1.697
GRPO-style, anchored $\beta = 1.5$, 60	0.010	11.2	1.806
GRPO-style, anchored $\beta = 1.0$, 120	0.047	3.1	1.835

and 36.3 words per response. At step 60 the values are 0.281, 0.069, and 81.6 words. These development-only measurements illustrate fewer errors accompanying reduced participation, but do not isolate the penalty from token-level credit or training duration. They motivate monitoring omission and length alongside error counts; raw-output recovery is still needed for independent rescoring.

5 Limitations and Future Work

Limitations. Our work carries several limitations. First, the corpus is built upon English-language earnings calls of publicly listed US firms, which follow Reg-FD rules. The framework and empirical findings may not directly transfer to markets with distinct securities regimes or non-English investor-relations discourse. Second, our rule-based numeric verifier targets quantitative claims, and cannot fully detect subtle qualitative misstatements or soft factual misrepresentations beyond numeric figures. Third, we test our framework on static source materials provided alongside each analyst question; we do not handle dynamic real-time document updates or multi-turn, long-horizon question–answer dialogues.

Future Work. There are several promising directions for future research. First, we plan to extend the corpus and evaluation benchmark to multilingual earnings calls and diverse regulatory environments to examine the generalizability of our decoupled alignment principle. Second, we will augment the verification module to cover qualitative factual claims, combining rule-based checks with lightweight, verifiable fact-checking components. Third, we aim to adapt DAF to multi-turn IR dialogues, where prior conversation history shapes the compliance constraints and managerial response style. Finally, the decoupled optimization paradigm can be transferred to other high-stakes regulated generative tasks, such as clinical or legal response generation, where style and factual grounding must be jointly preserved.

6 Conclusion

This work addresses managerial response generation for earnings-call question answering, where disclosure consistency, numerical grounding and authentic managerial register are inherently conflicting objectives. We adopt a decoupled optimization principle and introduce IR-Agent, a three-module Decoupled Alignment Framework that separates style learning, numeric behavioral tuning and candidate verification. By keeping style signals out of reward optimization and relying on a rule-based numeric verifier, IR-Agent delivers zero verifier-detected fabrication on 573 evaluation questions with a 4B-parameter base model, while preserving managerial linguistic style. Our closed-form, LLM-judge-free evaluation protocol and company-split corpus further support rigorous benchmarking for this task. Empirical ablations reveal how reward weighting and candidate-pool coverage govern the achievable operating point. We hope this decoupled paradigm offers a practical blueprint for high-stakes regulated generative tasks requiring both stylistic fidelity and factual rigor.

AI USE STATEMENT

Generative AI tools assisted research discussions, code preparation and inspection, literature discovery, and manuscript drafting, and were used to prepare the architecture illustration. Every numerical result reported in this paper was recomputed from saved model outputs by deterministic scripts, and selected references were checked against primary sources; the training and selection descriptions were checked against the available code. The authors take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.662/.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Brian J. Bushee, Ian D. Gow, and Daniel J. Taylor. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121, 2018. doi: 10.1111/1475-679X.12179. URL https://doi.org/10.1111/1475-679X.12179.

Richard A. Cazier, Jeff L. McMullin, and John S. Treu. Are lengthy and boilerplate risk factor disclosures inadequate? an examination of judicial and regulatory assessments of risk factor language. *The Accounting Review*, 96(4):131–155, 2021. doi: 10.2308/TAR-2018-0657. URL https://doi.org/10.2308/TAR-2018-0657.

Huizhong Chen and Huan Zhang. Lie to me: Detecting managerial evasiveness in earnings calls via conversational audio encoders. *arXiv preprint arXiv:2609.13893*, 2026. URL https://arxiv.org/abs/2609.13893.

Zhiyu Chen et al. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://aclanthology.org/2021.emnlp-main.300/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Richard Frankel, Matthew Johnson, and Douglas J. Skinner. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37(1), 1999. doi: 10.2307/2491400. URL https://doi.org/10.2307/2491400.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a. URL https://arxiv.org/abs/2210.10760.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://arxiv.org/abs/2305.14627.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. *arXiv preprint arXiv:2410.01039*, 2024. URL https://arxiv.org/abs/2410.01039.

Ian D. Gow, David F. Larcker, and Anastasia A. Zakolyukina. Non-answers during conference calls. *Journal of Accounting Research*, 59(4):1349–1384, 2021. doi: 10.1111/1475-679X.12371. URL https://doi.org/10.1111/1475-679X.12371.

Stephan Hollander, Maarten Pronk, and Erik Roelofsen. Does silence speak? an empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research*, 48(3):531–563, 2010. doi: 10.1111/j.1475-679X.2010.00365.x. URL https://doi.org/10.1111/j.1475-679X.2010.00365.x.

Bengt Holmström and Paul Milgrom. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24–52, 1991. URL https://academic.oup.com/jleo/article-abstract/7/special_issue/24/2194011.

Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Co-trained retriever-generator framework for question generation in earnings calls. *arXiv preprint arXiv:2409.18677*, 2024. URL https://arxiv.org/abs/2409.18677.

Mirae Kim, Seonghun Jeong, and Youngjun Kwak. A shaky voice is not always a dodge: Benchmarking textual and vocal evasion detection in earnings calls. *arXiv preprint arXiv:2608.28040*, 2026. URL https://arxiv.org/abs/2608.28040.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://arxiv.org/abs/2109.07958.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x. URL https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Shijian Ma, Yan Lin, and Yi Yang. EvasionBench: A large-scale benchmark for detecting managerial evasion in earnings call Q&A. *arXiv preprint arXiv:2601.09142v2*, 2026. URL https://arxiv.org/abs/2601.09142v2.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.14251.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://aclanthology.org/2022.emnlp-main.748/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Huzaifa Pardawala, Siddhant Sukhani, Agam Shah, Veer Kejriwal, Abhishek Pillai, Rohan Bhasin, Andrew DiBiasio, Tarun Mandapati, Dhruv Adha, and Sudheer Chava. SubjECTive-QA: Measuring subjectivity in earnings call transcripts' QA through six-dimensional feature analysis. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.20651.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 2023. URL https://arxiv.org/abs/2112.12870.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.13085.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Fengbin Zhu et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. URL https://aclanthology.org/2021.acl-long.254/.

A Data and Recomputation Details

A.1 Dataset construction and provenance

The construction code selects companies before selecting instances. From a recorded candidate pool of 135,354 question–answer pairs, training examples are filtered for numeric-string support, stratified by question type and answer length, and limited per company. The 4,000-example intermediate set becomes 3,895 after length filtering and 2,252 after retaining numeric answers. Our local audit independently recovers 2,252 such examples from the saved length-filtered file, covering 186 companies. Validation has 260 examples from 13 companies, and test has 575 from 20; these final counts should not be confused with earlier 600/588-question versions or an older 14-company validation split. Company intersections are empty between the final train, validation, and test sets. The archived test file matches the numerical-answerability criteria for every row.

Fact retrieval uses structured company facts and the call's reporting period. Prepared remarks are sentence-ranked by inverse-document-frequency-weighted overlap with the analyst question and financial-item terms, after the construction script's eligibility filtering; up to 40 sentences are returned in source order. Reference answers are used to accept or reject instances, not as retrieval queries or runtime inputs. The construction consequently selects an answerable subset; it does not establish performance on questions whose materials lack an answer. Source timestamps, revisions, transcript permissions, and pretraining exposure remain separate issues.

A.2 Available configuration and missing metadata

Table 5 reports the settings that could be recovered from the experiment record and the metadata that remains missing. Script defaults are not silently substituted for missing run metadata.

Table 5: Configuration provenance. Script defaults are not silently substituted for missing run metadata. The reported scores are from saved candidate outputs.

Component	Setting and provenance
Base	Qwen3-4B-Instruct-2507.
SFT starting point	`adapters_v4/checkpoint-15`; 2,252 examples; the research record states numeric-token weight 8. The exact weighted trainer, numeric mask, normalization, launch arguments, and final adapter configuration must be recovered together.
Adaptation	LoRA; the archived generic trainer targets attention and MLP projections. Its default rank 16 and alpha 32 are not sufficient evidence of the final v4 run settings.
Rollout pool	1,000 training prompts, eight answers each; fixed during updates. Sampling script: temperature 0.8, top-p 0.95, 300 new tokens.
Disclosure update	RLOO advantages; linear reward $g - 2f$; 60 updates, batch 8, learning rate 10^{-5}, AdamW, gradient norm clip 1.0; answer-token mean NLL.
Inference	Temperature 0.8, top-p 0.95, 300 new tokens; $N = 8$; filter $f = 0$, rank g, otherwise minimize f.
Long anchored arm	Anchored $\beta = 1.0$, GRPO-style; evaluated file traced to `rl_grpo_anchor_1.0_s240/checkpoint-120`. This differs in duration from the 60-step arms.
Export	Earlier quantised artifact is not confirmed to be the reported RLOO checkpoint; configuration reconciliation or re-export is required.

A.3 Metric implementation

Numerical extraction operates on distinct strings after comma and leading-zero removal, keeping at least two digits. It does not preserve signs as part of the string, convert percentages and currency scales, normalize decimal representations, or associate values with entities. Financial years and other multi-digit quantities can be counted. A response without an extracted numeric string has $g = f = 0$, so a clean response is not necessarily informative.

Precision is micro-aggregated, $\sum_i g_i / \max(\sum_i (g_i + f_i), 1)$; unsupported and supported counts are macro-averaged per question. Omission conditions on the human reference having at least one supported number. All common test examples meet that condition. Usable rate counts pools with any zero-unsupported candidate. Pool-average cleanliness counts individual responses across the first eight candidates, and should not be confused with the first-candidate $N = 1$ cleanliness.

Nominalization uses a case-insensitive suffix pattern (including `tion`, `sion`, `ment`, `ance`, `ence`, `ity`, and `ness`) divided by the English word-token count and multiplied by 100. Deflection uses a fixed phrase pattern and sentence splitting at punctuation; this can mis-segment decimals. Copying is the percentage of distinct, case-sensitive answer word 4-grams also present in the materials. It is neither a sentence-copying rate nor evidence of plagiarism. Mechanism-language and active-construction measures are binary pattern indicators, not semantic explanation or grammatical-voice judgements. Words use whitespace tokenization. The lexical repetition proxy is one minus the unique-token fraction, which is length-sensitive; empty output is fewer than 20 stripped characters.

Metric formulas. Let $\mathcal{C}_i^{(n)}$ be the first n generated candidates for question i, I the evaluated questions, $Q \subseteq I$ those whose human reference cites at least one supported number, and $\mathbf{1}[\cdot]$ the indicator. The reported quantities are

$$\begin{aligned}
\text{prec} &= \frac{\sum_i g_i}{\max\left(\sum_i (g_i + f_i),\, 1\right)}, & \text{miss} &= \frac{1}{|Q|} \sum_{i \in Q} \mathbf{1}\left[g(a_i^\star) = 0\right], \\
\text{use} &= \frac{1}{|I|} \sum_{i \in I} \mathbf{1}\left[\exists\, a \in \mathcal{C}_i^{(n)} :\ f(a) = 0\right], & \text{cln} &= \frac{\sum_i \sum_{a \in \mathcal{C}_i^{(n)}} \mathbf{1}[f(a) = 0]}{\sum_i \left|\mathcal{C}_i^{(n)}\right|}, \qquad (7) \\
\text{nom}_{100}(a) &= 100\, \frac{\left|S_{\text{nom}}(a)\right|}{W(a)}, & \text{defl}(a) &= \frac{\left|S_{\text{defl}}(a)\right|}{\left|S(a)\right|}, \\
\text{copy}(a) &= 100\, \frac{\left|G_4(a) \cap G_4(F)\right|}{\left|G_4(a)\right|}, & \text{rep}(a) &= 1 - \frac{\left|U(a)\right|}{\left|T(a)\right|},
\end{aligned}$$

where $W(a)$ counts word tokens, $S_{\text{nom}}(a)$ and $S_{\text{defl}}(a)$ the answers' sentences matching the suffix and deferral-phrase patterns, $G_4(a)$ the set of distinct case-sensitive word 4-grams, and $U(a)$ and $T(a)$ the unique and total token counts. Mechanism-explanation and active-voice rates are binary pattern indicators, $\text{emp}(a) = 100\, \mathbf{1}\left[|\text{strip}(a)| < 20\right]$, and every per-answer quantity is reported as its mean over I unless stated otherwise.

Every table is rescored with `metrics_full.py`, avoiding the different style-pattern versions embedded in earlier generation scripts. Determinism applies to scoring saved text, not to regeneration or API reproducibility. No seed-averaged uncertainty or significance test is claimed.

A.4 Shared system instruction

The recorded instruction is reproduced to specify the role and evidence boundary:

> You are an investor-relations assistant helping a company answer analyst questions on an earnings call. Base your answer on the facts provided below, and do NOT invent any number that is not in those facts. If a specific detail is not covered by the facts, give a natural, graceful response such as 'we don't break that out' or 'we'll provide more color in future quarters' — never reply with 'no information available'. Do not make forward-looking commitments without hedging words like 'subject to' or 'may'. Answer directly and clearly.

The instruction itself offers phrases detected by the deflection metric. Deflection therefore requires interpretation against question answerability and human reference values; it is not an unconditional penalty on justified refusals.

A.5 Validation: SFT alone versus SFT plus selection

Table 6 is independently rescored from the archived 260-question validation outputs. Recomputed supported and unsupported counts match all 2,080 stored candidates for each policy. SFT alone has

Table 6: Validation-set results (260 questions). Single response and selected response are different conditions; the RLOO test omission advantage over SFT plus selection does not carry over to validation.

Configuration	N	Fab.	Miss (%)	All-dirty pools
Base alone	1	0.142	30.4	28/260
Base + selection	8	0.069	17.3	16/260
SFT alone	1	0.362	60.4	59/260
SFT + selection	8	0.004	16.9	1/260
SFT+RLOO alone	1	0.315	56.9	55/260
SFT+RLOO + selection	8	0.000	20.8	0/260

60.4% omission; selection reduces it to 16.9%. RLOO plus selection removes the one remaining all-dirty pool but has higher omission (20.8%). Thus, the test-set omission gain from adding RLOO is not a uniform advantage across splits. The full system's conditional cleanliness and its participation level must be discussed separately.

A.6 Aggregate view of the main comparison

Figure 2 shows the two headline quantities together, sorted by fabrication, with the human reference and the full system marked.

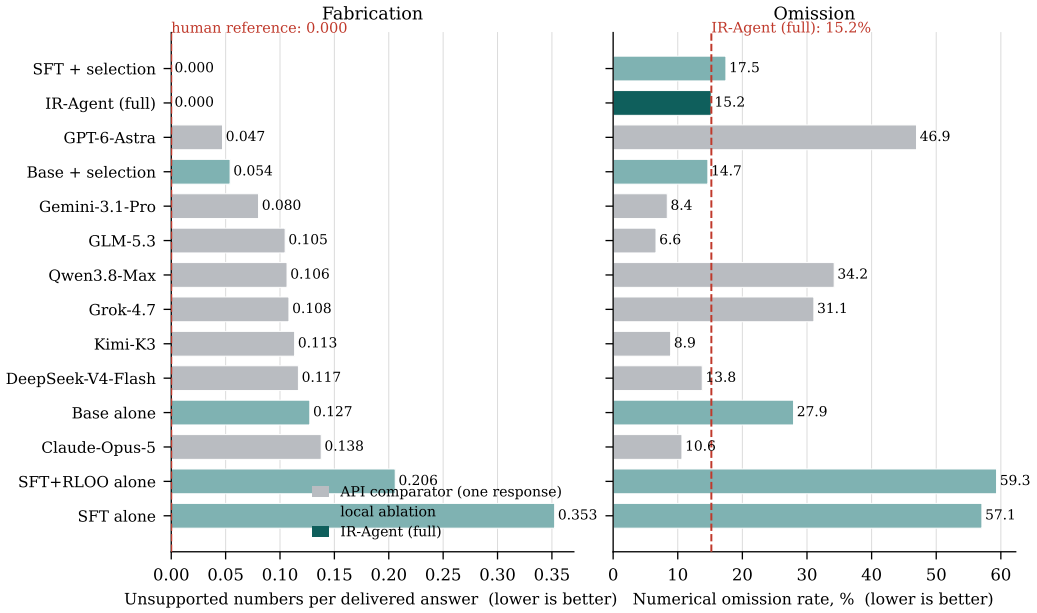

Figure 2: **Fabrication and omission for every compared system** on the 573 common questions, sorted by fabrication. Grey bars are API comparators (one response each), light teal bars are local ablations, and the dark teal bar is the full system. Dashed lines mark the human reference (0.000 unsupported numbers) and the full system (15.2% omission); lower is better in both panels. Five comparators omit less than the full system, while GPT-6-Astra and Qwen3.8-Max omit about three and two times as much. The API system with the lowest fabrication also cites the fewest disclosed numbers (1.251 against the human reference 1.417).

A.7 Additional reward and development controls

The code additionally implements $r_{\text{ratio}} = g^2/(g+f)$, defined as zero for $g+f=0$. No test result is assigned to it without the corresponding archived output. RFT selects the highest-reward

candidate in each training group and applies supervised learning with unit weight; it is not a fourth policy-gradient estimator. Table 7 lists the recorded controls.

Table 7: Development-only controls transcribed from the experiment record. They are not pooled with the recomputed test results; raw-candidate recovery and common rescoring remain outstanding.

Configuration	Evaluation	Unsupported	Supported	Nom.	Words
RFT, step 15	Valid, $N = 8$	0.000	1.423	1.461	—
RFT, step 60	Valid, $N = 8$	0.000	1.608	1.602	—
High penalty + token credit, 60	Single response	0.069	0.281	—	81.6
High penalty + token credit, 120	Single response	0.023	0.085	—	36.3

The recorded RFT pools contain at least one clean candidate for all 260 validation questions. The high-penalty measurements are for the variant with added token-level numerical credit, not an isolated penalty-only experiment. They illustrate reduced numerical participation alongside reduced unsupported counts, without establishing a general impossibility of learning faithfulness.

A.8 One question, four systems and the human reference

We quote one test question (FORECAST type) with the answers it received. Among the 573 evaluation questions, this is the only one on which the quoted systems jointly show an unsupported number, an answer containing no number at all, and heavy source overlap. The materials supplied with the question contain 30 distinct numeric strings. Table 8 gives the verifier's measurements for each quoted answer.

Analyst question. *"Kathy, you highlighted a few moments ago that free cash flow growth is the real story here. So I want to focus on that since you mentioned it. Can you talk about multiyear free cash flow growth? You've done so in the past. But how do you think about the biggest drivers from a top line networking, net income drop-through, tax, CapEx perspective?"*

Human reference — 4 grounded, 0 unsupported, 1.1% copying

"I'm happy to provide some additional color there. The 2024 updates that we provided on the call today should give you a sense for our near-term sales and earnings growth expectations underlying that free cash flow expansion. [...] And so we expect that capital intensity to be alleviating in 2025 and beyond. [...] So all in all, a number of key tailwinds most critical of which, of course, is the expectation of continued growth and margin opportunity in the business."

GPT-6-Astra — 0 grounded, 0 unsupported, 3.4% copying

"We see multiyear free cash flow growth coming from several factors working together. Top-line growth and the conversion of our backlog should support operating earnings, while continued program execution, productivity, and working-capital discipline should improve cash conversion. [...] Taxes should also become a more meaningful tailwind over time as current tax and pension-related cash-flow headwinds may reverse, although the timing remains subject to the tax profile of the business [...] Taken together, we see a path to meaningful cash-flow expansion through operating growth, better conversion, tax normalization, and disciplined investment."

The answer never states a figure.

Claude-Opus-5 — 11 grounded, 1 unsupported, 17.7% copying

"This quarter we grew revenue 9% year-over-year, and we've raised our full-year 2023 guidance by $800 million throughout the year to approximately $39 billion. [...] Operating income this quarter was up over 20% year-over-year to just over a billion dollars, with a 10.4% margin. Net income of $937 million grew over 2% [...] We generated nearly $900 million in free cash flow this quarter and remain on track to hit our 2023 target of $2.25 billion to $2.65 billion."

Its single flagged string is "937", which the materials carry as "937.00": a representation mismatch rather than an invented figure.

Gemini-3.1-Pro — 9 grounded, 0 unsupported, 37.5% copying

"We remain very confident in the long-term value creation opportunity from our free cash flow expansion. We currently see a path to grow our cash flows at a greater than 20% CAGR next year and through 2025, with further expansion anticipated in the following years. [...] CapEx: We anticipate CapEx to be roughly consistent as a percentage of sales in 2024 before declining in 2025 and beyond. Taxes and Pensions: We expect that pension and tax cash flow headwinds will reverse over the next few years. [...] This helped generate nearly $900 million in free cash flow in the quarter on $937.00 million of reported net income."

IR-Agent (full) — 5 grounded, 0 unsupported, 0.9% copying

"Yes. I think, Chris, our free cash flow guidance is still in the $2.25 billion to $2.65 billion range for 2024. And so you can still look at that as a very solid year-to-year growth. And we're adding to that 2024 guidance, as we did earlier today. We're adding about $400 million in sales growth and we're adding about $400 million in operating income. And so, that's really a full-year 2024 sales growth of 4% to 5% and operating income of 4% to 5%. And again, as you know, when you add in the pension income and tax adjustments, this will be a solid year-over-year cash flow growth of over 20%. So, we're really looking at that as a very strong performance for the year."

Table 8: Verifier measurements for the five answers quoted above, computed with the same rule used throughout the paper. Copies of these answers are retained with the evaluation artifacts.

Answer	Grounded	Unsupported	Copy (%)	Words
Human reference	4	0	1.1	367
GPT-6-Astra	0	0	3.4	208
Claude-Opus-5	11	1	17.7	328
Gemini-3.1-Pro	9	0	37.5	267
IR-Agent (full)	5	0	0.9	125

The two frontier answers that avoid unsupported numbers do so in opposite directions: one states no figure, the other reproduces its source material. Both satisfy the checker and neither answers the question. The full system cites five disclosed figures in 125 words, with copying at 0.9% against the human reference's 1.1%.

A.9 Recomputation artifacts

The local writing package includes a recomputation script, its metric JSON, a dataset audit, validation scores, and the author-selected figure. Main and ablation tables are generated directly from these recomputed summaries. Responses are sorted by stored candidate index and the first N are used, with stable candidate-order tie breaking. The source material, raw experimental outputs, and earlier drafts are retained separately; a completed public release is not asserted.