Corpus-Derived Style Templates Improve LLM Financial Research Q&A

Bryan Healey
Aiera

July 6, 2026

Abstract

Conversational assistants for institutional finance are expected to answer in the register of professional sell-side research: conclusion-first, figures woven into prose, calibrated hedging, field-standard vocabulary and analytical framings. We study whether such a register can be *distilled automatically from a corpus* and injected as a system-prompt *style template* to improve a chat model's research answers. From ∼12,800 research documents across all providers—spanning equities, credit, commodities, and macro/FX—we extract per-document style and domain guidance and synthesize a single **global template**. In a blinded, open-book A/B evaluation, the global template beats a prose-capable no-template baseline by a large margin under a Claude judge (**90.5%** pairwise win-rate with Opus 4.8 answering, 95% CI $[82.3, 95.1]$, $p < 0.001$; **62.6%** with Haiku 4.5, pooled over three runs, $[56.4, 68.4]$, $p < 0.001$). The result is robustly cross-vendor: three independent non-Anthropic judges (Mistral, Amazon Nova, DeepSeek) favor the template in every condition (79–92% on the stronger model's answers), with no consistent same-vendor inflation. Template use also *raises* LLM-judged key-fact recall (+2.6/+3.9 points), consistent with substantive rather than purely cosmetic gains. A second experiment finds that **per-topic specialization does not help**: across 89 topic templates and a ∼190-item, 30-topic test set, topic templates do not beat the global template even under oracle routing (38.9% win-rate on the stronger model, a statistical tie on the weaker; 0/30 topics significant after FDR correction).

1 Introduction

Style and format strongly shape the perceived and actual quality of LLM answers, yet most deployments rely on hand-written style instructions that are labor-intensive and rarely validated. We ask three questions: **(Q1)** can a useful style template be *derived automatically* from a corpus of expert writing? **(Q2)** does it improve answers on real research Q&A, and is any gain *substantive* (more complete/faithful) or merely *stylistic*? **(Q3)** does *topic-specific* specialization beat a single global template? We answer Q1–Q2 affirmatively and Q3 negatively, each with blinded pairwise evaluation, multiple judge models (including cross-vendor), and a key-fact faithfulness metric.

2 Related work

Our setup connects three threads. **LLM-as-judge** evaluation is now standard for open-ended generation, with strong models matching human preference at high agreement (>80%, comparable to human–human agreement; 1) and structured rubric/chain-of-thought judging improving alignment with human ratings [2]. Such judges exhibit known position [3] and verbosity [1] biases, and a tendency to favor LLM-generated text [2]; we mitigate these by blinding, randomizing answer order (following 3), and reporting two judges plus an independent (separately prompted) recall metric.

Prompt-level behavior steering—e.g. governing model outputs with a written set of principles [4]—shows prose behavior is highly steerable from the prompt; we differ in *deriving* the guidance from a domain corpus rather than authoring it, which connects to a broader line of corpus-derived style prompting and to the general question of specialized versus monolithic prompts (prompt routing).

Retrieval and grounding [5]: we evaluate open-book, holding facts constant, which separates the contribution of *style guidance* from that of *evidence*—most prior style work conflates the two. Finally, domain LLMs for finance such as Wu et al. [6] pursue capability via pretraining; we instead target answer *quality and register* via a lightweight, corpus-derived prompt. To our knowledge, the global-vs-per-topic trade-off for corpus-derived *style* templates in grounded financial Q&A has not been directly studied.

3 Template construction

All language models in this work—for template construction, answer generation, and judging, including the cross-vendor judges of §5.1—are accessed through **Amazon Bedrock** via its Converse API, so every model (Anthropic, Mistral, Amazon, and DeepSeek) is served under one provider with comparable inference settings.

Corpus. ∼12,800 documents sampled from all available providers, restricted to 2024-onward and spanning ∼12 geographic regions. Sampling is provider-balanced and asset-class-diverse: an initial equity-heavy draw was supplemented by targeted collection of prose-bearing non-equity research (sovereign and high-yield credit, commodities, FX/macro), giving a final mix of roughly two-thirds equity and one-third non-equity.

Per-document analysis (Claude Sonnet 4.6). Each document yields a canonical topic, region, and a TEMPLATE—natural-language guidance on prose style: sentence construction, data integration, tone/hedging, technical-vocabulary handling, comparison/trend phrasing, transitions, and uncertainty expression. A domain layer adds reusable KEY TERMS, FORMULAS (general metric definitions, not document-specific values), and RELATIONSHIPS (how data interconnects—drivers, decompositions, framing comparisons).

Synthesis (Claude Opus 4.8). Per-document templates are merged hierarchically (batches → merge) into one global template (v15): 8 style sections plus consolidated *Key Terms*, *Key Formulas*, and *Analytical Relationships* (11 total)—the domain sections now span equity, credit, rates/FX, and commodity conventions. Topic labels were normalized to a canonical vocabulary (sector in the topic; geography as a separate facet), giving 89 topics with ≥10 documents, each synthesized into a per-topic template for Experiment 2.

4 Evaluation protocol

Task — grounded (open-book) Q&A. Each item is a question plus the source excerpt(s) containing the answer. Both arms receive identical context, so the only manipulated variable is the style template; this isolates the template's effect and enables faithfulness scoring against the relevant ground truth.

Arms. *Baseline*: a neutral but prose-capable system prompt (explicitly told to write well-structured prose, not bullet lists)—so we measure *answer quality*, not a prose-vs-bullets formatting artifact. *Templated*: the same base prompt plus the rendered template.

Metrics. (i) Blinded pairwise preference, judged by two models (Opus 4.8, Sonnet 4.6) with randomized answer order (to counter position bias; 3), asked which answer better *answers the question* (accuracy, completeness, clarity). We report win-rate over *decisive* (non-tie) pairs with Wilson 95% CIs and two-sided sign-test p-values. (ii) A 1–5 rubric (completeness, accuracy, directness, clarity, professional tone). (iii) Faithfulness: the fraction of the item's atomic `key_facts` stated correctly, scored by the primary judge in a separate call—a substance-oriented metric independent of the style rubric, though (like the rubric) it is itself an LLM judgment, not a deterministic match.

Test sets. *Experiment 1*: 100 items stratified from the private corpus (question + gold answer + key_facts + sources), rebalanced across sectors. *Experiment 2*: ∼190 items generated with Opus 4.8 from held-out corpus documents, covering the 30 highest-volume topics (5–8 each)—now spanning equities, credit, commodities, and macro/FX—each grounded with gold answer, key_facts, and source excerpt.

5 Experiment 1 — does the template help?

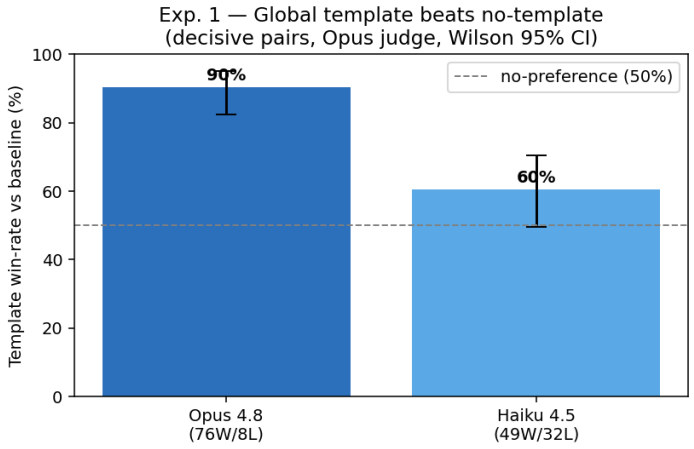

Figure 1: Global template win-rate vs. no-template baseline (decisive pairs, Opus judge, Wilson 95% CIs). The template wins decisively on both model tiers.

The template wins over non-template on both model tiers and both judges (Table 1)—rather decisively for Opus; the Haiku win margin is smaller, and we establish its significance by pooling repeated runs (§5.1). Win-rates are over decisive pairs; CIs are Wilson 95%; p-values are two-sided sign tests.

Table 1: Templated vs. baseline (pairwise).

Answering model	Judge	Win-rate	95% CI	W/L	p
Opus 4.8	Opus	**90.5%**	[82.3, 95.1]	76 / 8	5.0e−15
Opus 4.8	Sonnet	89.5%	—	51 / 6	—
Haiku 4.5	Opus	**60.5%**	[49.6, 70.4]	49 / 32	7.5e−02
Haiku 4.5	Sonnet	82.9%	—	58 / 12	—

The gains appear substantive, not merely cosmetic. LLM-judged faithfulness *rises* under the template—Opus 92.2% → 94.7% (+2.6), Haiku 85.0% → 89.0% (+3.9; Figure 2)—so style guidance at least does not trade accuracy for polish. The rubric localizes the lift to *completeness* (+0.22 Opus / +0.36 Haiku) and *professional tone* (+0.30 / +0.32), with accuracy and directness roughly flat. These faithfulness and rubric deltas are small and near ceiling, and we do not test them for significance (only the pairwise win-rate is); we therefore read them as *corroborating* the pairwise result rather than as standalone evidence. Wins are broad across question types (Figure 3; Opus: cross-doc-synthesis 100%, thesis 100%, guidance-change 88%, KPI-extraction 88%, materiality 80%).

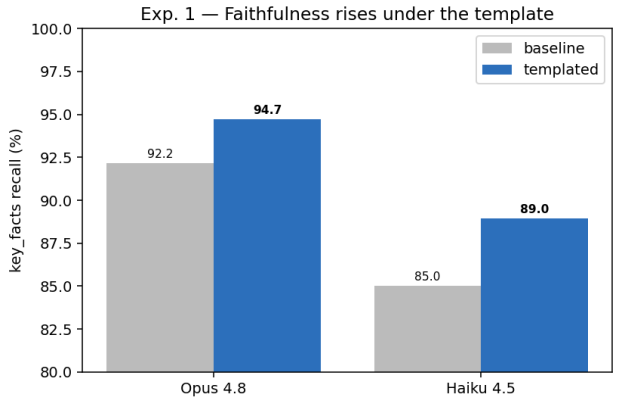

Figure 2: LLM-judged `key_facts` recall rises under the template on both models.

Where it helps least. On the weaker model (Haiku) the gains concentrate in completeness and tone; clarity is roughly flat (−0.05) and simple KPI-extraction is a coin-flip (50% win-rate), where analytical elaboration adds little over a crisp factual lookup. These remain the natural refinement targets.

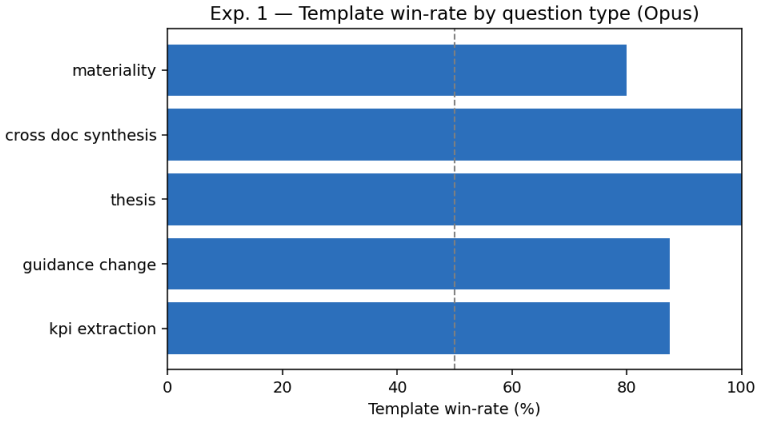

Figure 3: Template win-rate by question type (Opus answering model, Opus judge).

5.1 Robustness: run variance and cross-vendor judging

Two checks probe whether the Experiment-1 result is an artifact of (a) single-run noise or (b) same-vendor judging.

Run variance (Haiku). The Haiku cell is the weakest headline, so we ran it three times. Across runs the win-rate is 60.5%, 59.0%, 67.8% (Opus judge; SD ≈4); single runs are underpowered (p ranges 0.001–0.14). Pooling all three gives **62.6%** (154/92 decisive, n=246, 95% CI $[56.4, 68.4]$, $p < 0.001$), which we adopt as the Haiku headline; the effect is robust but more modest than on Opus, and single-run p-values understate variance.

Cross-vendor judging. To test self-preference (the answer generator and primary judges are all Claude; LLM judges can favor LLM text, 2), we re-scored the *same* stored Experiment-1 answers with three non-Anthropic judges via Bedrock: Mistral Large, Amazon Nova Pro, and DeepSeek R1 (a different-lineage reasoning model). All four judges, on both answer sets, favor the template (Figure 4)—the advantage is robustly cross-vendor. We see *no consistent same-vendor inflation*: on Opus-generated answers the Claude judge (90.2%) sits mid-pack among the neutral judges (78.9–92.5%, DeepSeek highest), and on Haiku-generated answers the Claude judge is the *most conservative* (61.0% vs. the neutral judges' 80.8–87.9%). If anything, the neutral judges rate the template's benefit at least as highly as the same-vendor judge. Judge–judge agreement was modest (Claude vs. each neutral judge, 51–72%).

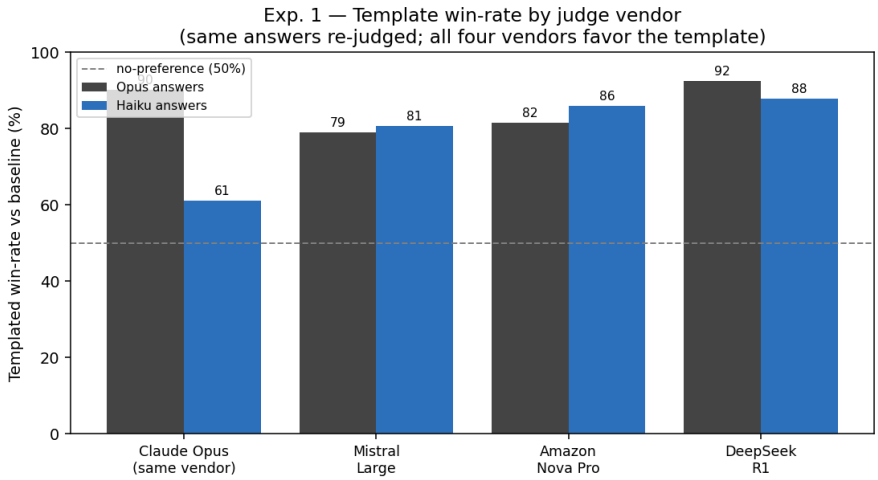

Figure 4: Experiment-1 template win-rate when the *same* answers are re-judged by four judge vendors. All bars exceed 50% on both answer sets; the same-vendor Claude judge is mid-pack on Opus answers and the most conservative on Haiku answers.

6 Experiment 2 — does per-topic specialization help?

We compared three arms (baseline / global / topic), selecting the topic template by *oracle* (the item's true topic—an upper bound that removes routing error).

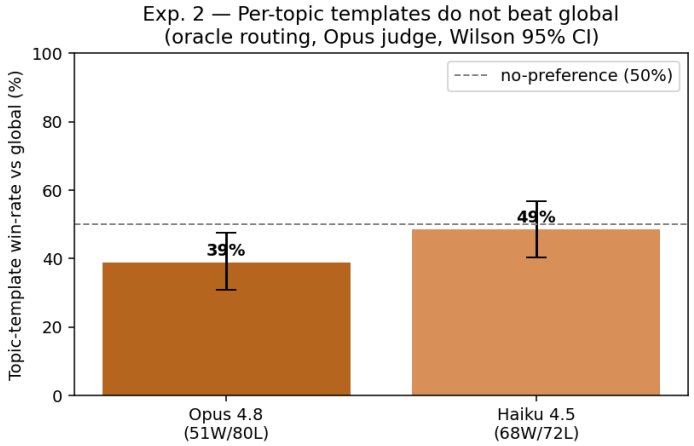

Figure 5: Per-topic template win-rate vs. the global template (oracle routing, Opus judge, Wilson 95% CIs). Neither bar exceeds the 50% no-preference line: the stronger model clearly favors global, the weaker model is at parity.

The test set covers the 30 highest-volume topics (i.e. the topics best resourced for building a specialized template), now spanning equities, credit, commodities, and macro/FX—so the broader corpus is exercised, which if anything favors the topic arm. Per-topic templates nonetheless do not beat the global template, even with perfect routing: on the stronger

Table 2: Topic vs. global (oracle routing, Opus judge).

Model	Topic w.r.	95% CI	W/L	p	Glob. v. base	Router
Opus 4.8	**38.9%**	[31.0, 47.5]	51 / 80	1.4e−02	85.1%	71%
Haiku 4.5	**48.6%**	[40.4, 56.8]	68 / 72	8.0e−01	70.6%	72%

model they significantly lose (38.9%, $p = 0.014$), and on the weaker model they are statistically indistinguishable from global (48.6%, $p = 0.80$). Faithfulness was a wash (Opus global 55.7% vs. topic 55.1%; Haiku 49.4% vs. 49.3%). Of 30 per-topic comparisons, **0 were significant after Benjamini–Hochberg FDR correction** (q=0.05)—and the newly added non-equity topics (credit, commodities, macro/FX) do not favor specialization either. Meanwhile global-vs-baseline replicated strongly on this set (85% / 71%), corroborating Experiment 1 on items independent of its test set (though, as in Exp. 1, generated and judged by Claude; see §8). Absolute faithfulness is lower here (44–56%) than in Exp. 1, reflecting harder, more specific questions; this floor applies equally to all three arms and does not affect the topic-vs-global comparison.

Interpretation. The global template is itself domain-enriched and synthesized from all ∼12,800 documents, making it broad and robust; each topic template draws on far fewer documents and is narrower. In open-book Q&A the source already supplies facts, so topic-specific scaffolding adds little over strong general guidance. A real router (≈71% accurate) would only erode the topic arm, which already fails to beat global at the oracle ceiling.

7 Conclusion

A single, corpus-derived global style template yields a large and statistically significant preference improvement on grounded research Q&A across a frontier and a small model, with faithfulness and rubric signals consistent with substantive (not merely cosmetic) gains, and is corroborated by three independent non-Anthropic judges (§5.1) that favor it at least as strongly as the same-vendor judge. Per-topic specialization does not beat the global template—clearly losing on the stronger model and at parity on the weaker—so we recommend deploying the single global template rather than building topic routing.

8 Limitations & future work

- **All-Claude pipeline (circularity), partially addressed.** The template is synthesized, the Exp. 2 items are authored, the answers are generated, and the primary judges are Claude; LLM judges can favor LLM text [2]. We directly tested this for Exp. 1 with three non-Anthropic judges (§5.1): the template's advantage holds across all of them (79–92% on the stronger model's answers), with no consistent same-vendor inflation. We did *not* cross-vendor re-judge Exp. 2 (the harness did not persist answer texts, so it would require regeneration), and human evaluation remains outstanding—both are the key next steps.

- **Mostly single-run.** The Haiku Exp. 1 cell was replicated three times and pooled (§5.1); all other cells are single non-replicated runs with non-deterministic generation/judging. Single-run Wilson CIs and p-values treat one run as the sample and

understate true variability, so small per-cell p-values are not evidence of stability; broader replication is future work.

- **"Faithfulness" is LLM-judged.** Our substance metric is scored by the primary judge, not by deterministic matching against `key_facts`; it is independent of the style rubric prompt but shares the judge's biases. Treat the faithfulness and rubric deltas as corroborating, not as an independent ground-truth check.
- **Open-book only.** Holding facts constant isolates style; closed-book effects (on recall/hedging) are untested, and most corpus content is proprietary so closed-book correctness is not measurable here.
- **Refinement.** The weak-model gains are thin on simple KPI-extraction (a concise mode for direct lookups may help); and we have not yet tested whether the domain layer pays off more in closed-book or tool-augmented settings.

References

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL `https://arxiv.org/abs/2306.05685`. arXiv:2306.05685.

[2] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL `https://arxiv.org/abs/2303.16634`. arXiv:2303.16634.

[3] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL `https://arxiv.org/abs/2305.17926`. arXiv:2305.17926.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL `https://arxiv.org/abs/2212.08073`.

[5] 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://arxiv.org/abs/2005.11401`. arXiv:2005.11401.

[6] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023. URL `https://arxiv.org/abs/2303.17564`.