Analogical Deep Research: Retrieving and Integrating Historical Analogies for Foresight Analysis

Anonymous Authors[1]

Abstract

Historical analogies – past events that share a mechanism with the current situation, are among the most powerful tools for foresight analysis. We introduce *Analogical Deep Research* (ADR), a task that asks LLM agents to retrieve and integrate such analogies, and the first benchmark ADR-bench covering 15 historical and forward events. We find that state-of-the-art deep research (DR) agents largely fail at ADR, defaulting to surface matching. We argue ADR is inherently a causal problem and prove that (i) surface-level methods are non-identifiable for hidden mechanisms, while (ii) two or more cross-confirming analogies are sufficient to distinguish structural necessity from coincidence. Then, we propose **C**ausal **Ana**logical Researcher (CANA), an agent that decomposes events into general structural forms and synthesizes a Structural Analogy Brief for analogy generation and integration. CANA improves analogy generation by up to 10% over the state of the art and substantially outperforms commercial DR agents on ADR-bench.

1. Introduction

Systematic comparisons between current situations and structurally similar past events—*historical analogies*—are widely used for foresight by policymakers (Neustadt & May, 1986), scientists (Dunbar, 1995), and practitioners (Dalio, 2017; Keulen, 2023), with Green & Armstrong (2007) reporting that structurally aligned analogies raise expert conflict-prediction accuracy from 32% to 60%. Untrained humans, however, tend to settle for superficial parallels (Ghilani et al., 2017; Khong, 1992).

In parallel, large language models (Bubeck et al., 2025; Guo et al., 2025) and *deep research* (DR) agents that combine planning with external tools (OpenAI, 2025; Li et al., 2025b; Team et al., 2026) now solve expert-level question answering (Phan et al., 2025; Wei et al., 2025; Mitchener et al., 2025), search under complex constraints, and even research-style synthesis. This raises the question:

Can DR agents do analogical search and reasoning for foresight?

Early evidence is mixed: LLMs struggle with story analogies (Jiayang et al., 2023), cross-domain parallels (Sourati et al., 2024), long-context retrieval (Ye et al., 2024), and exhibit human-like biases to surface similarity (Li et al., 2025a).

We introduce Analogical Deep Research (ADR), the task of aligning a target event with mechanism-matched historical events for foresight analysis, and build the first benchmark ADR-bench of 15 events: 10 historical events with literature-documented analogies and 5 unfolding events without ground-truth analogies, spanning finance, geopolitics and technology. On ADR-bench, four state-of-the-art commercial DR agents (OpenAI DR, Gemini DR, Qwen DR, MiroFlow) rarely use historical analogies at all, and even when they do, the resulting foresight is shallow and lacks structural grounding.

ADR is inherently *causal*: it requires uncovering the hidden mechanism shared across events (Keulen, 2023). We prove that surface matching is non-identifiable for hidden positions, whereas two or more cross-confirming analogies suffice to recover structural necessity. Motivated by this, we propose **C**ausal **Ana**logical Researcher (CANA), which decomposes events into preconditions–temporal chains–mechanisms–outcomes, iteratively retrieves analogies that jointly cover the target's hidden positions, and synthesizes a Structural Analogy Brief (SAB) for downstream DR.

Contributions. (i) We formalize ADR and release ADR-bench, showing existing DR agents do not retrieve or integrate historical analogies; (ii) we derive two principles—*mechanism alignment* and *cross-analogy confirmation*—from a causal-identifiability analysis; (iii) we propose CANA, the first ADR agent, achieving up to 10% improvement on

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

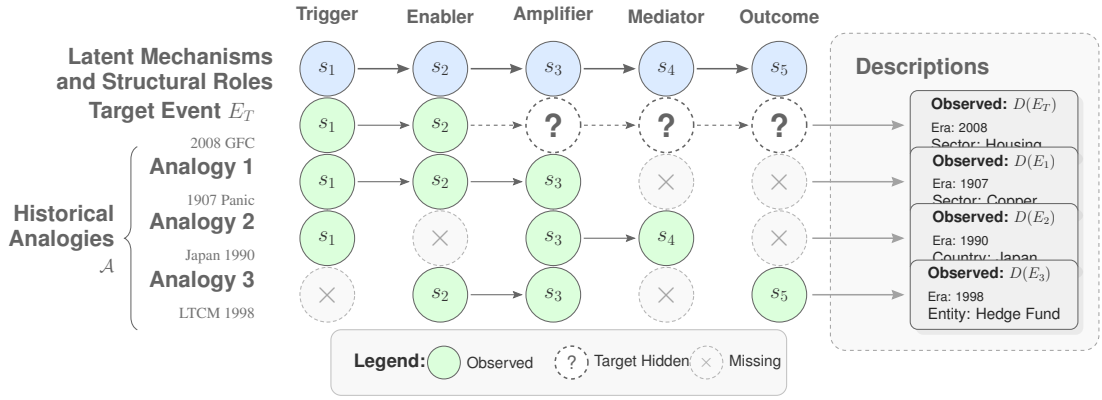

Figure 1. Analogical Deep Research. The target E_T (2008 GFC at Aug. 2007) reveals only s_1–s_2 while s_3–s_5 are hidden. Each historical analogy (1907 Panic, Japan 1990, LTCM 1998) covers a different subset of positions; their union covers all hidden positions, and the amplifier s_3 recurs across all three (interlocks, keiretsu, OTC exposure)—a cross-analogy invariant suggesting a hidden amplifier (CDS web) in the target. Surface descriptions differ substantially, making retrieval hard for DR agents.

analogy generation and large gains over commercial DR agents on ADR-bench.

2. Foresight with Historical Analogies

Setup. An event E has a *descriptive* representation $D(E) \in \mathbb{R}^p$ (entities, domain, timeline) and a *mechanistic* one $M(E)=(V_E, \mathcal{E}_E)$ encoding causal factors and relations, with $D(E)=h(M(E), \chi_E)$ for event-specific context χ_E (Gentner, 1983; Clement & Gentner, 1991a). An analogy E_S for target E_T shares structural positions $\{s_1, \ldots, s_L\}$ in the underlying causal graph, discovered through cross-event alignment rather than fixed in advance. At cutoff t_c only part of V_{E_T} is observed, so positions split into observed $\mathcal{S}_T^{\text{obs}}$ and hidden $\mathcal{S}_T^{\text{hid}}$. The foresight risk is $\mathcal{R}(\mathcal{A}, w)=\sum_s \mu_s \, \mathrm{TV}(\widehat{P}_s^{\mathcal{A}}, P_s^w)$ with $\widehat{P}_s^{\mathcal{A}}$ the forecast under analogy set $\mathcal{A}$. Under *mechanism transfer* (Keulen, 2023): if E_S, E_T share position s and E_S has progressed beyond E_T, $\mathrm{TV}(\widehat{P}_s^{E_S}, P_s^T) \leq \alpha_s$—a mechanism that recurs across matched events transfers up to bounded error.

Theorem 1 (Surface non-identifiability)**.** *If two admissible worlds w^0, w^1 share full surface observations $\mathcal{O}_D^\infty(w^0)=\mathcal{O}_D^\infty(w^1)$ but differ on whether a hidden position s is active ($Z_s(w^0) \neq Z_s(w^1)$), then no surface-level method can identify Z_s: $\inf_{\hat{Z}_s} \max_j \Pr_{w^j}(\hat{Z}_s \neq Z_s(w^j)) \geq 1/2$. If the worlds imply $\mathrm{TV}(P_F^{w^0}, P_F^{w^1}) \geq 2\Delta_s$, any surface forecast incurs $\max_j \mathrm{TV}(\widehat{P}_F, P_F^{w^j}) \geq \Delta_s$, regardless of how many analogies are retrieved.*

Theorem 1 (proof in App. C) explains why LLMs matching on $D(E)$ fail to recover hidden mechanisms (Sourati et al., 2024; Li et al., 2025a). Given partial observability a single analogy may align coincidentally, but cross-confirmation across analogies rules out chance:

Theorem 2 (Cross-analogy confirmation)**.** *Let $X_{k,s}=\mathbb{1}[E_k \text{ confirms } s]$ with $P(X_{k,s}=1 \mid Z_s=1)=q_{k,s}$, $P(X_{k,s}=1 \mid Z_s=0)=p_{k,s}$, independent across k given Z_s. Then the posterior log-odds equal $\log \frac{\pi_s}{1-\pi_s} + \sum_{k\in\mathcal{A}}[X_{k,s} \log \frac{q_{k,s}}{p_{k,s}} + (1-X_{k,s}) \log \frac{1-q_{k,s}}{1-p_{k,s}}]$, so each independent confirming analogy multiplies the odds by $q_{k,s}/p_{k,s}$ and coincidence decays exponentially.*

For the calibration ($\pi_s=0.5, q_s=1, p_s=0.2, \delta=0.05$), two independent analogies per position suffice—paralleling causal discovery from multiple datasets (Huang et al., 2020) and multi-view causal representation learning (Yao et al., 2024; Brehmer et al., 2022).

Two design principles. P1 (Thm. 1): retrieve on $M(E)$, not $D(E)$. **P2** (Thm. 2): require $K_s^\star \geq 2$ independent analogies per hidden position.

ADR-bench. ADR-bench consists of 15 events: 10 *historical* events with literature-documented analogies and 5 *forward* unfolding events, spanning finance, geopolitics and technology. Each is annotated with cutoff, pre-cutoff brief, causal structure, oracle analogies, cross-analogy invariants and hidden factors. Agents see only the pre-cutoff brief and a neutral prompt. Evaluation uses a 5-level claim rubric (L1 surface fact → L4 cross-analogy hidden-factor inference) plus analysis quality (RAS), foresight depth (FQS) and mechanism grounding (Gr); see App. D.

Preliminary findings. Fig. 2 evaluates OpenAI DR, Gemini DR, Qwen DR and MiroFlow on ADR-bench: all four rarely cite analogies, and even when they do, the analysis lacks structural grounding—motivating the design of CANA below.

3. Causal Analogical Researcher

CANA realizes the two principles via a structural decomposition and a reflective generator integrated with an open-

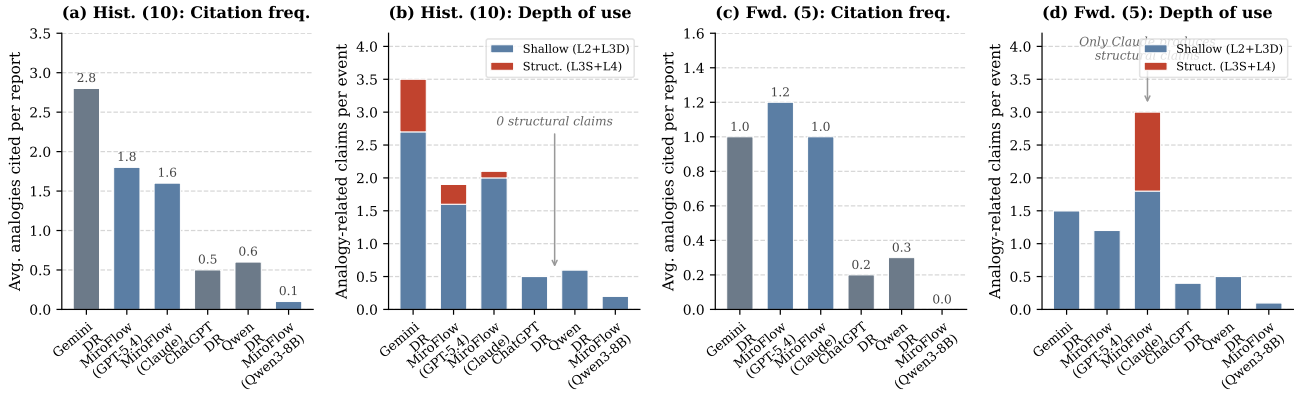

Figure 2. Preliminary results: existing commercial DR agents barely use historical analogies and produce shallow foresight on ADR-bench.

Table 1. Detailed evaluation results on General Analogies using a 9-item structured rubric. The column headers denote: **C. Isom.** (Chain Isomorphism), **Dir. Con.** (Direction Consistency), **Map. Con.** (Mapping Consistency), **Over. Suf.** (Overlap Sufficiency), **Idio. Cov.** (Idiosyncrasy Coverage), **Sys. Alig.** (System Alignment), **Nov.** (Novelty), **Surf. Sim.** (Surface Similarity), and **Calib. B.** (Calibration Bonus). Final represents the weighted composite score. All evaluations are conducted by GPT-5.4. Best results within each LLM block are **bolded**, and second best are underlined.

Method	C. Isom.	Dir. Con.	Map. Con.	Over. Suf.	Idio. Cov.	Sys. Alig.	Nov.	Surf. Sim.	Calib. B.	Final
gpt-5.4-mini										
Direct Gen.	1.46	1.77	1.68	1.84	0.96	0.95	**2.07**	2.02	-0.69	3.31
Summarizing	**1.96**	**2.16**	**2.26**	2.75	1.52	1.79	1.47	2.77	-0.65	4.49
Self-reflection	1.94	2.10	2.23	2.64	**1.57**	**1.88**	1.18	2.77	-0.58	4.12
CANA	1.95	2.14	2.20	**2.77**	1.54	1.78	1.84	2.64	-0.57	**5.22**
sonnet-4-6										
Direct Gen.	1.69	2.07	1.96	2.07	1.29	1.48	**2.99**	2.19	-0.46	3.43
Summarizing	1.89	2.05	2.21	2.66	1.41	1.68	1.58	2.71	-0.67	4.65
Self-reflection	1.97	**2.15**	**2.24**	2.74	1.50	1.88	1.73	2.73	-0.56	5.23
CANA	**1.98**	2.11	2.21	**2.84**	**1.52**	**1.91**	1.89	2.62	-0.47	**5.70**
qwen3-8b										
Direct Gen.	0.855	1.18	0.59	0.26	0.11	0.10	**2.94**	0.36	-0.17	2.14
Summarizing	**1.88**	**2.08**	2.11	2.53	1.39	1.64	1.79	2.64	-0.66	4.37
Self-reflection	1.87	2.04	**2.11**	**2.66**	**1.43**	**1.66**	1.73	2.69	-0.68	4.63
CANA	1.77	1.94	1.98	2.66	1.34	1.49	2.06	2.54	-0.70	**4.87**

source DR framework.

Structural decomposition. Following Clement & Gentner (1991b), we prompt the LLM to decompose every event into `preconditions`, `temporal chains`, `mechanisms`, `outcomes`—a generic instantiation of trigger/enabler/amplifier/mediator/outcome roles.

Structural reflective generation. The LLM proposes candidates from the decomposed target; feedback flags positions uncovered by the current set (P2), and the LLM iterates until each hidden position is confirmed by ≥ 2 analogies.

Structural Analogy Brief. CANA plugs into MiroFlow (Team et al., 2026): the SAB contains mechanism statements, per-analogy limitations, cross-analogy invariants and predicted hidden factors, consumed by the main DR agent. Unlike Li et al. (2025a) who reflect on descriptive dimensions, our reflection is structurally grounded and enforces cross-confirmation.

4. Experiments

Historical analogy generation. Following Li et al. (2025a), we compare direct generation, summary-then-generate, self-reflection and CANA on Qwen3-8B, GPT-5.4-mini and Sonnet-4.6. Table 1 (9-dim structural rubric) shows CANA wins on all backbones with up to 25% Final-score gain over self-reflection ($\sim$ 10% MDS gain; App. F); Qwen3-8B+CANA surpasses GPT-5.4-mini+self-reflection, indicating the lift

Table 2. Results on `ADR-bench`. Left: 10 historical events (22 hidden factors). Right: 5 forward events (20 hidden factors, outcomes unfolding). L3-S+L4 = cross-analogy structural claims per event; HF@L4 = hidden factors inferred via analogical reasoning; RAS = analysis quality; FQS_d = depth-filtered foresight quality; Gr% = fraction of claims with mechanism grounding ≥ 2. All commercial `DR` agents produce **zero** HF@L4 across both splits.

		Historical events (10)					Forward events (5)				
	Agent	**L3S+L4**	**HF@L4**	**RAS**	**FQS_d**	**Gr%**	**L3S+L4**	**HF@L4**	**RAS**	**FQS_d**	**Gr%**
Comm.	Gemini DR	0.8	0/22	0.967	0.71	0.18	0.0	0/20	0.889	0.79	0.21
	ChatGPT DR	0.0	0/22	0.741	0.53	0.13	0.0	0/20	0.630	0.47	0.07
	Qwen DR	0.0	0/22	0.689	0.37	0.15	0.0	0/20	0.600	0.27	0.08
Vanilla	`MiroFlow` (GPT-5.4)	0.3	1/22	0.933	0.73	0.34	0.0	0/20	0.837	0.86	0.42
	`MiroFlow` (Sonnet-4.5)	0.1	0/22	0.819	0.82	0.35	1.2	1/20	0.800	0.55	0.12
	`MiroFlow` (Qwen3-8B)	0.0	0/22	0.319	0.29	0.07	0.0	0/20	0.570	0.00	0.00
Ours	`CANA` (Sonnet-4.5)	**14.0**	**13/22**	**1.000**	0.77	0.74	**16.6**	3/20	**1.000**	0.56	0.84
	`CANA` (GPT-5.4)	11.6	7/22	0.993	**0.92**	**0.94**	13.4	3/20	**1.000**	**0.86**	**0.94**
	`CANA` (Qwen3-8B)	3.2	1/22	0.607	0.82	0.81	6.0	**3/20**	0.807	0.62	0.81

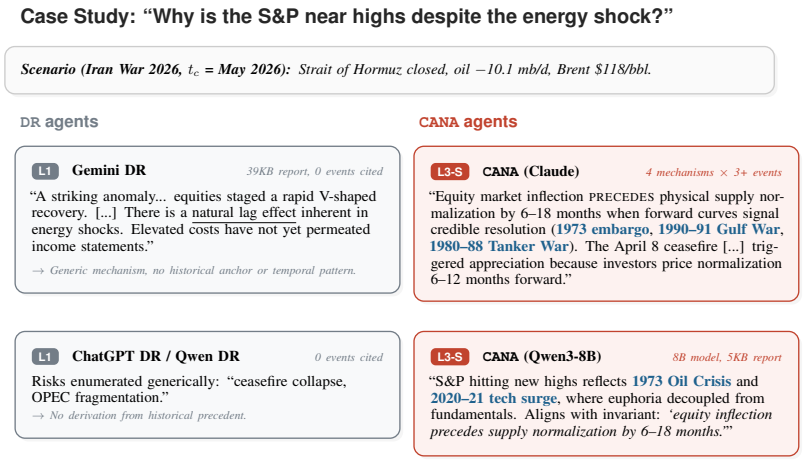

Figure 3. How different agents handle the same analytical question (Iran War 2026). **Left** (gray): baseline agents produce L1 claims regardless of report length or model capability. **Right** (red): `CANA` agents produce L3-S claims grounded in $\geq$2 historical events.

is method-driven.

ADR setup. We evaluate 9 agents on `ADR-bench`: 3 commercial `DR` (`OpenAI DR`, `Gemini DR`, `Qwen DR`), 3 vanilla `MiroFlow` and 3 `CANA` variants (`Sonnet-4.5`, `GPT-5.4`, `Qwen3-8B`); identical prompts; `Sonnet-4.5` as judge.

Findings (Table 2). **(F1)** Commercial `DR` agents score 0/22 HF@L4 on historical and 0/20 on forward events—confirming Thm. 1. **(F2)** `CANA` lifts HF@L4 to 13/22 and L3-S+L4 to 14.0/event; `Qwen3-8B+CANA` surpasses every commercial agent. **(F3)** `Sonnet-4.5` maximizes claim/HF volume; `GPT-5.4` maximizes grounding (Gr= 0.94) and depth (FQS_d=0.92) at RAS= 1.00—validating Thm. 2.

Case study. Fig. 3: agents reconcile S&P 500 at new highs with elevated oil prices during the Iran War 2026. Baselines cite a generic "lag effect" (L1); `CANA` retrieves the 1973/1979/1990–91 oil shocks and produces an L3-S claim grounded in three mechanism-aligned analogies ("equity inflection PRECEDES supply normalization by 6–18 months").

5. Conclusion

We introduced Analogical Deep Research, derived two causal-identifiability principles for foresight from analogies, built the first benchmark `ADR-bench`, and proposed `CANA`, the first `ADR` agent. `CANA` improves analogy retrieval by $\sim$10% and substantially outperforms commercial `DR` agents on `ADR-bench`, with case studies confirming the qualitative gain.

References

Achenbaum, W. A. The making of an applied historian: Stage two. *The Public Historian*, 5(2):21–46, 1983.

Adams, J., Hansen, N., and Zhang, K. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bartha, P. F. A. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press, 2010.

Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Brunk, D. C. Curing the Somalia syndrome: Analogy, foreign policy decision making, and the Rwandan genocide. *Foreign Policy Analysis*, 4(3):301–320, 2008.

Bubeck, S., Coester, C., Eldan, R., Gowers, T., Lee, Y. T., Lupsasca, A., Sawhney, M., Scherrer, R., Sellke, M., Spears, B. K., Unutmaz, D., Weil, K., Yin, S., and Zhivotovskiy, N. Early science acceleration experiments with gpt-5. *ArXiv*, abs/2511.16072, 2025.

Clement, C. A. and Gentner, D. Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15 (1):89–132, 1991a.

Clement, C. A. and Gentner, D. Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15 (1):89–132, 1991b.

Dalio, R. *Principles: Life and Work*. Simon & Schuster, New York, 2017. ISBN 9781501124020.

Dunbar, K. How scientists really reason: Scientific reasoning in real-world laboratories. In Sternberg, R. J. and Davidson, J. E. (eds.), *The Nature of Insight*, pp. 365–395. MIT Press, 1995.

Falkenhainer, B., Forbus, K. D., and Gentner, D. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, 1989.

Gentner, D. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

Ghilani, D., Luminet, O., Erb, H.-P., Flassbeck, C., Rosoux, V., Tames, I., and Klein, O. Looking forward to the past: An interdisciplinary discussion on the use of historical analogies and their effects. *Memory Studies*, 10(3):274–285, 2017.

Green, K. C. and Armstrong, J. S. Structured analogies for forecasting. *International Journal of Forecasting*, 23(3): 365–376, 2007.

Guldi, J. and Armitage, D. *The history manifesto*. Cambridge University Press, 2014.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Houghton, D. P. The role of analogical reasoning in novel foreign-policy situations. *British Journal of Political Science*, 26(4):523–552, 1996.

Huang, B., Zhang, K., Gong, M., and Glymour, C. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of AAAI*, 2020.

Jiayang, C., Qiu, L., Chan, T. H., Fang, T., Wang, W., Chan, C., Ru, D., Guo, Q., Zhang, H., Song, Y., Zhang, Y., and Zhang, Z. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of EMNLP*, pp. 11518–11537, 2023.

Karger, E., Bastani, H., Chen, Y.-H., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *Proceedings of ICLR*, 2025.

Keulen, S. Historical analogies: Functions, limitations and the correct use of historical analogies in applied history. *Journal of Applied History*, 5(2):111 – 131, 2023. doi: 10.1163/25895893-bja10036. URL https://brill.com/view/journals/joah/5/2/article-p111_2.xml.

Khong, Y. F. *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965*. Princeton University Press, 1992.

Li, N., Yuan, S., Chen, J., Liang, J., Wei, F., Liang, Z., Yang, D., and Xiao, Y. Past meets present: Creating historical analogy with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3942–3957, 2025a.

Li, X., Jin, J., Dong, G., Qian, H., Zhu, Y., Wu, Y., Wen, J.-R., and Dou, Z. WebThinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025b.

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., and Scialom, T. GAIA: A benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983*, 2023.

Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., Landsness, E. C., Barabási, D. L., Narayanan, S., Evans, N., Reddy, S., Foiani, M. S., Kamal, A., Shriver, L. P., Cao, F., Wassie, A. T., Laurent, J. M., Melville-Green, E., Ramos, M. C., Bou, A., Roberts, K. F., Zagorac, S., Orr, T. C., Orr, M. E., Zwezdaryk, K. J., Ghareeb, A. E., McCoy, L., Gomes, B., Ashley, E. A., Duff, K. E., Buonassisi, T., Rainforth, T., Bateman, R. J., Skarlinski, M., Rodriques, S. G., Hinks, M. M., and White, A. D. Kosmos: An ai scientist for autonomous discovery. *ArXiv*, abs/2511.02824, 2025.

Nersessian, N. J. *Creating Scientific Concepts*. MIT Press, 2008.

Neustadt, R. E. and May, E. R. *Thinking in Time: The Uses of History for Decision-Makers*. Free Press, 1986.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, February 2025. Accessed: 2026-05-07.

Opiełka, G., Rosenbusch, H., and Stevenson, C. E. Analogical reasoning inside large language models: Concept vectors and the limits of abstraction. *arXiv preprint arXiv:2503.03666*, 2025.

Parsons, M. and Nalau, J. Historical analogies as tools in understanding transformation. *Global Environmental Change*, 38:82–96, 2016.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Sen, P., Workiewicz, M., and Puranam, P. Can LLMs aid analogical reasoning for strategic decisions? a comparative study. *Strategy Science*, 11:118–136, 2026.

Sourati, Z., Ilievski, F., Sommerauer, P., and Jiang, Y. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086, 2024.

Team, M. et al. Miroflow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv preprint arXiv:2602.22808*, 2026.

Webb, T., Holyoak, K. J., and Lu, H. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541, 2023.

Wei, J. et al. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Xu, R. and Peng, J. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*, 2025.

Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., and Locatello, F. Multi-view causal representation learning with partial observability. In *Proceedings of ICLR*, 2024.

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E. H., and Zhou, D. Large language models as analogical reasoners. In *Proceedings of ICLR*, 2024.

Ye, X., Wang, A., Choi, J., Lu, Y., Sharma, S., Shen, L., Tiyyala, V. M., Andrews, N., and Khashabi, D. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of EMNLP*, 2024.

Yu, J., He, R., and Ying, R. Thought propagation: An analogical approach to complex reasoning with large language models. In *Proceedings of ICLR*, 2024.

Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J. W., and Khashanah, K. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*, 2023.

Yuan, S., Chen, J., Sun, C., Liang, J., Xiao, Y., and Yang, D. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of ACL (Long Papers)*, pp. 1249–1265, 2024.

Zeng, Z., Liu, J., Chen, S., He, T., Liao, Y., Tian, Y., wangjinpeng.levi, Wang, Z., YangYang, Yin, L., Yin, M., Zhenwei, Z., Cai, T., Chen, X., Chen, Z., Chen, J., Du, Y., Gao, X., Guo, J., HU, L., Jiao, J., Li, X., Liu, J., nishuang, Wen, Z., Zhang, G., Zhang, K., xin zhou, Blanchet, J., Qiu, X., Wang, M., and Huang, W. Futurex: An advanced live benchmark for LLM agents in future prediction. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhang, W., Zhao, L., Xia, H., Sun, S., Sun, J., Qin, M., Li, X., Zhao, Y., Zhao, Y., Cai, X., et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*, 2024.

A. Related Work

Historical analogies for foresight analysis. In human history, historical analogies are widely used or cited for foresight (Achenbaum, 1983; Guldi & Armitage, 2014; Parsons & Nalau, 2016; Ghilani et al., 2017; Keulen, 2023). For example, policymakers often exploit analogies to reason about unprecedented scenarios (Neustadt & May, 1986; Houghton, 1996; Khong, 1992; Brunk, 2008). Scientists use analogical reasoning to identify useful hypotheses (Dunbar, 1995; Nersessian, 2008). Green & Armstrong (2007) also showed that forecasting with analogies can be largely improved from 32% to 60%. Analogies refer to events that share similar mechanisms or aligned structures (Gentner, 1983; Falkenhainer et al., 1989). Structural alignment of analogies helps with foresight analysis (Bartha, 2010), which is inherently a causal problem where separate partial observations are aligned to recover the underlying causal relations (Huang et al., 2020; Adams et al., 2021; Yao et al., 2024). We exploit the conceptual connection between analogy identification and causal learning to build our framework.

Analogical reasoning with LLMs. Due to the usefulness of analogies, there is a growing body of work using LLMs for analogical reasoning (Webb et al., 2023). Opiełka et al. (2025) showed LLMs can abstract concepts and perform analogical reasoning. Yasunaga et al. (2024) found that self-generated analogous examples help improve LLM reasoning. Yu et al. (2024) showed that relating analogical sub-problems helps solve a challenging composite problem with LLMs. As LLMs are trained on extensive datasets, finding analogies with LLMs is also of great interest (Jiayang et al., 2023; Yuan et al., 2024; Sourati et al., 2024; Li et al., 2025a). Jiayang et al. (2023) found that LLMs are limited in finding analogies from stories. Sourati et al. (2024) showed limitations in finding cross-domain analogies. Ye et al. (2024) provided evidence on the limitations of LLMs in finding analogies in long contexts. The most relevant work is Li et al. (2025a), which introduced a historical-analogy benchmark and showed that reflecting on topics, background, process, and result can improve historical analogy generation. None of these works fully define the problem and the principles required for historical analogy identification and integration.

Deep research and forecasting with LLMs. Deep research (DR) agents systematically equip LLMs with planning and tool use (Xu & Peng, 2025) for complex real-world problems requiring multi-step reasoning (OpenAI, 2025; Li et al., 2025b). For example, Li et al. (2025b); Team et al. (2026) showed that synthesis of comprehensive search results helps tackle challenging reasoning tasks (Wei et al., 2025; Mialon et al., 2023; Phan et al., 2025) and even future prediction (Zeng et al., 2026). Forecasting is inherently challenging (Karger et al., 2025) but valuable downstream (Yu et al., 2023; Zhang et al., 2024; Sen et al., 2026). Despite the importance of historical analogies, we find that DR agents do not proactively exploit them, and propose the first ADR agent.

LLM Use Statement

From the research side, this work studies the use of LLMs for analogical deep research. LLMs are also used to help do evaluation and experimentation. From the paper writing side, we use LLMs to assist with improving the writing of this work.

Limitation and Future Works

As a pilot work in ADR, the scale of ADR-bench might be a bit limited, and the evaluation also relies on the LLMs. Future works could scale up the curation of ADR-bench as well as more fine-grained and human-aligned evaluation protocols.

Broader impacts

We study using LLMs for Analogical Deep Research that will benefit the whole humanity and society. This work does not involve human subjects or personally identifiable information beyond public benchmarks used under their licenses.

Table 3. Notation for the ADR theory.

Symbol	Meaning
E_T	Target event
$\mathcal{U}$	Candidate analogy universe
$D(E)$	Descriptive representation of event E
$M(E)$	Mechanistic representation of event E
$\mathfrak{P}$	Abstract causal pattern
$\mathcal{S}_{\mathfrak{P}}$	Structural positions in pattern $\mathfrak{P}$
ψ_E	Instantiation map from event factors to structural positions
ϕ_k	Structural mapping from analogy E_k to target E_T
Ω_{pos}	Target-side position-level overlap
$\mathcal{O}_D^\infty$	Full surface observation
$\mathcal{O}_M^\infty$	Full mechanism-level observation
Z_s	Indicator that position s is truly active / structural
$X_{k,s}$	Indicator that analogy E_k confirms position s
$\widehat{P}_s^{\mathcal{A}}$	Analogy-based forecast for position s
P_s^w	True trajectory distribution at position s in world w
P_F^w	True foresight distribution in world w
α_s^{tr}	Mechanism-transfer error bound
Δ_s	Position-wise two-world separation gap
Δ_F	Foresight-distribution two-world separation gap
δ_s	Posterior probability that inferred position s is false

B. Additional Technical Details

B.1. Notation

C. More Details of the Theory

C.1. Problem Definition

Definition 3 (Event Representation)**.** An event E has two representations: (i) **descriptive** $D(E) \in \mathbb{R}^p$, capturing observable surface features (entities, domain, timeline); and (ii) **mechanistic** $M(E) = (V_E, G_E, \{K_{v,E}\}, \theta_E)$, where $V_E = \{v_1, \ldots, v_k\}$ are causal factors, $G_E = (V_E, \mathcal{E}_E)$ is a directed graph encoding causal relationships, $K_{v,E}$ are transition kernels, and θ_E are event-specific parameters. Each factor $v \in V_E$ has an associated trajectory $\tau_v : [0, T] \to \mathbb{R}$. The descriptive and mechanistic representations are related by a many-to-one mapping $D(E) = h(M(E), \chi_E)$, where χ_E is event-specific context.

Definition 4 (Abstract Pattern and Structural Position)**.** An abstract causal pattern is $\mathfrak{P} = (\mathcal{S}_{\mathfrak{P}}, G_{\mathfrak{P}})$, where $\mathcal{S}_{\mathfrak{P}} = \{s_1, \ldots, s_L\}$ is a finite set of structural positions and $G_{\mathfrak{P}}$ is a directed graph over $\mathcal{S}_{\mathfrak{P}}$. Event E **instantiates** $\mathfrak{P}$ via a graph homomorphism $\psi_E : V_E \to \mathcal{S}_E \subseteq \mathcal{S}_{\mathfrak{P}}$, preserving edge structure. Factors $v \in V_{E_1}$ and $w \in V_{E_2}$ occupy the **same structural position** if $\psi_{E_1}(v) = \psi_{E_2}(w)$. Positions are not pre-specified but discovered through cross-event alignment, following Gentner's structure-mapping theory (Gentner, 1983): objects are placed in correspondence according to their roles in the common relational structure (Clement & Gentner, 1991a).

Definition 5 (Structural Mapping and Position-Level Overlap)**.** A structural mapping from source E_S to target E_T is a partial injection $\phi : V_{E_S} \rightharpoonup V_{E_T}$ satisfying edge preservation and position consistency: $\psi_{E_S}(v) = \psi_{E_T}(\phi(v))$ for all $v \in \text{dom}(\phi)$. The **position-level overlap** for analogy set $\mathcal{A}$ is $\Omega_{\text{pos}}(\mathcal{A}, E_T) = |\bigcup_k \psi_{E_T}(\text{im}(\phi_k))| \,/\, |\mathcal{S}_T|$, measured on the target side.

Definition 6 (Partial Observation and Hidden Factors)**.** At cutoff t_c, the analyst observes $V^{\text{obs}} \subseteq V_{E_T}$ and

$$\mathcal{S}_T^{\text{obs}} = \psi_{E_T}(V^{\text{obs}}).$$

The true active position set is

$$\mathcal{S}_T^{\text{true}} = \psi_{E_T}(V_{E_T}).$$

Hidden factors are

$$\mathcal{V}_T^{\text{hid}} = V_{E_T} \setminus V^{\text{obs}},$$

and hidden positions are

$$\mathcal{S}_T^{\text{hid}} = \mathcal{S}_T^{\text{true}} \setminus \mathcal{S}_T^{\text{obs}}.$$

Definition 7 (ADR Problem)**.** Given target E_T at cutoff t_c with partial observation, and a universe $\mathcal{U}$ of historical events with fully observed representations, the ADR problem is to select $\mathcal{A}^* \subseteq \mathcal{U}$ and produce foresight report $\mathcal{F} = (\hat{\mathcal{V}}^{\text{hid}}, \hat{Y}_T, \mathcal{C}, \mathcal{K})$ maximizing:

$$\mathcal{A}^* = \arg \max_{|\mathcal{A}| \leq K_{\max}} \text{Qual}(\mathcal{F}(\mathcal{A})) - \text{Qual}(\mathcal{F}(\emptyset)) \tag{1}$$

where $\text{Qual}(\mathcal{F}) = \lambda_{\text{cov}} \cdot \text{Cov}_{\text{eval}} + \lambda_{\text{prec}} \cdot \text{Prec}_{\text{eval}} + \lambda_{\text{depth}} \cdot \text{Depth}_{\text{eval}}$.

Definition 8 (Surface Observation and Foresight Risk)**.** A **possible world** $w = (E_T, \mathcal{U}, \mathfrak{P}, \{\psi_E\}, t_c, \chi)$ specifies the target, analogy universe, pattern, instantiations, cutoff, and context.

For two-world comparisons, we evaluate positions on a common finite candidate universe $\mathcal{S}_\star \subseteq \mathcal{S}_{\mathfrak{P}}$. For each $s \in \mathcal{S}_\star$, define

$$Z_s(w) = \mathbb{1}[s \in \mathcal{S}_T^{\text{true}}(w)].$$

If s is inactive in world w, P_s^w denotes the corresponding null or baseline trajectory distribution.

The **full surface observation** is

$$\mathcal{O}_D^\infty(w) = (D(E_T), \{D(E_k)\}_{E_k \in \mathcal{U}}, \{\text{sim}(D(E_k), D(E_T))\}_{E_k \in \mathcal{U}}).$$

A **surface-level method** is any function of $\mathcal{O}_D^\infty(w)$.

The **full mechanism-level observation** is

$$\mathcal{O}_M^\infty(w) = \left(M^{\text{obs}}(E_T), \{M(E_k), \psi_{E_k}, \phi_k\}_{E_k \in \mathcal{U}}\right).$$

A **mechanism-level method** is any function of $\mathcal{O}_M^\infty(w)$.

For a structural position s, let τ_s denote the trajectory induced by the target factor(s) $v \in V_{E_T}$ with $\psi_{E_T}(v) = s$, or a fixed aggregation of those trajectories when multiple factors map to s. The **foresight target** is

$$Y_F = \Gamma(\{\tau_s(t) : s \in \mathcal{S}_\star, t > t_c\})$$

with true distribution P_F^w and forecast $\widehat{P}_F^{\mathcal{A}}$.

The **position-wise foresight risk** is:

$$\mathcal{R}(\mathcal{A}, w) = \sum_{s \in \mathcal{S}_\star} \mu_s \cdot \text{TV}(\widehat{P}_s^{\mathcal{A}}, P_s^w),$$

where $\mu_s \geq 0$, $\sum_s \mu_s = 1$.

C.2. Main Results

We state one assumption and two theorems. Theorem 10 is an information-theoretic impossibility result for surface matching; Theorem 13 quantifies the benefit of cross-analogy mechanism matching. For readability, we state the results for deterministic methods. Randomized methods satisfy the same bounds in expectation over their internal randomness.

Assumption 9 (Mechanism Transfer)**.** If factors $v_S \in V_{E_S}$ and $v_T \in V_{E_T}$ occupy the same structural position ($\psi_{E_S}(v_S) = \psi_{E_T}(v_T) = s$), both are active at their respective cutoffs, and E_S has progressed beyond E_T's current stage, then:

$$\text{TV}(\widehat{P}_s^{E_S}, P_s^T) \leq \alpha_s^{\text{tr}}$$

for some position-specific transfer-error bound $\alpha_s^{\text{tr}} > 0$.

Theorem 10 (Surface Non-Identifiability)**.** *Suppose there exist two admissible worlds w^0, w^1 such that $\mathcal{O}_D^\infty(w^0) = \mathcal{O}_D^\infty(w^1)$ (identical surface observations), but for a hidden position s: $Z_s(w^0) \neq Z_s(w^1)$, where $Z_s(w) = \mathbb{1}[s \in \mathcal{S}_T^{\text{true}}(w)]$. Then no surface-level method can identify whether s is active:*

$$\inf_{\hat{Z}_s = f(\mathcal{O}_D^\infty)} \max_{j \in \{0,1\}} \Pr_{w^j}(\hat{Z}_s \neq Z_s(w^j)) \geq \tfrac{1}{2} \tag{2}$$

If the two worlds imply separated foresight distributions, $\text{TV}(P_F^{w^0}, P_F^{w^1}) \geq 2\Delta_F$, then:

$$\inf_{\widehat{P}_F = f(\mathcal{O}_D^\infty)} \max_{j \in \{0,1\}} \text{TV}(\widehat{P}_F, P_F^{w^j}) \geq \Delta_F \tag{3}$$

This holds for all K including $K = \infty$: adding more surface-matched analogies cannot eliminate foresight risk caused by D-unidentifiable hidden mechanisms.

Proof. Let a possibly randomized surface-level estimator be written as

$$\hat{Z}_s = \varphi(\mathcal{O}_D^\infty, R),$$

where R denotes internal randomness independent of the world. Since

$$\mathcal{O}_D^\infty(w^0) = \mathcal{O}_D^\infty(w^1),$$

the distribution of $\hat{Z}_s$ is the same under w^0 and w^1.

Without loss of generality, suppose

$$Z_s(w^0) = 0, \qquad Z_s(w^1) = 1.$$

Let

$$r = \Pr(\hat{Z}_s = 1).$$

Because the estimator has the same output distribution in both worlds, its error probability in world w^0 is r, while its error probability in world w^1 is $1 - r$. Hence

$$\max_{j\in\{0,1\}} \Pr_{w^j}(\hat{Z}_s \neq Z_s(w^j)) = \max\{r, 1-r\} \geq \frac{1}{2}.$$

Taking the infimum over all surface-level estimators gives

$$\inf_{\hat{Z}_s = f(\mathcal{O}_D^\infty)} \max_{j\in\{0,1\}} \Pr_{w^j}(\hat{Z}_s \neq Z_s(w^j)) \geq \frac{1}{2}.$$

For the foresight part, let a possibly randomized forecast be

$$\widehat{P}_F = \Phi(\mathcal{O}_D^\infty, R).$$

Again, because the full surface observation is identical in w^0 and w^1, the forecast has the same distribution under both worlds. For any realization of R, the triangle inequality for total variation gives

$$\mathrm{TV}(P_F^{w^0}, P_F^{w^1}) \leq \mathrm{TV}(P_F^{w^0}, \widehat{P}_F) + \mathrm{TV}(\widehat{P}_F, P_F^{w^1}).$$

Taking expectation over R if the method is randomized, we obtain

$$\mathbb{E}_R\mathrm{TV}(P_F^{w^0}, \widehat{P}_F) + \mathbb{E}_R\mathrm{TV}(\widehat{P}_F, P_F^{w^1}) \geq \mathrm{TV}(P_F^{w^0}, P_F^{w^1}).$$

Therefore,

$$\max_{j\in\{0,1\}} \mathbb{E}_R\mathrm{TV}(\widehat{P}_F, P_F^{w^j}) \geq \frac{1}{2}\mathrm{TV}(P_F^{w^0}, P_F^{w^1}).$$

Using the assumed separation

$$\mathrm{TV}(P_F^{w^0}, P_F^{w^1}) \geq 2\Delta_F,$$

we get

$$\max_{j\in\{0,1\}} \mathbb{E}_R\mathrm{TV}(\widehat{P}_F, P_F^{w^j}) \geq \Delta_F.$$

For deterministic forecasts, the expectations over R disappear.

Finally, a finite-budget surface matcher with any budget K is only a measurable function of $\mathcal{O}_D^\infty$. Since even the full surface observation is identical in the two worlds, no finite subset of surface-matched analogies can distinguish them. Thus the lower bound holds for all finite K, and also when $K = \infty$. □

Lemma 11 (Position-wise Surface Lower Bound)**.** *Let $H \subseteq \mathcal{S}_\star$ be a set of candidate hidden positions. Suppose there exist two admissible worlds w^0, w^1 such that*

$$\mathcal{O}_D^\infty(w^0) = \mathcal{O}_D^\infty(w^1),$$

and for every $s \in H$,

$$\mathrm{TV}(P_s^{w^0}, P_s^{w^1}) \geq 2\Delta_s.$$

Define the restricted position-wise risk

$$\mathcal{R}_H(\mathcal{A}, w) = \sum_{s \in H} \mu_s \mathrm{TV}(\widehat{P}_s^{\mathcal{A}}, P_s^w).$$

Then every surface-level method satisfies

$$\inf_{\mathcal{A}_D} \max_{j \in \{0,1\}} \mathcal{R}_H(\mathcal{A}_D, w^j) \geq \sum_{s \in H} \mu_s \Delta_s.$$

Proof. Since $\mathcal{O}_D^\infty(w^0) = \mathcal{O}_D^\infty(w^1)$, any surface-level method produces the same position-wise forecast in both worlds. Hence, for every $s \in H$,

$$\widehat{P}_s^{\mathcal{A}_D}(w^0) = \widehat{P}_s^{\mathcal{A}_D}(w^1) = \widehat{P}_s^{\mathcal{A}_D}.$$

By the triangle inequality for total variation,

$$\mathrm{TV}(P_s^{w^0}, P_s^{w^1}) \leq \mathrm{TV}(P_s^{w^0}, \widehat{P}_s^{\mathcal{A}_D}) + \mathrm{TV}(\widehat{P}_s^{\mathcal{A}_D}, P_s^{w^1}).$$

Using the assumed separation

$$\mathrm{TV}(P_s^{w^0}, P_s^{w^1}) \geq 2\Delta_s,$$

we obtain

$$\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_D}, P_s^{w^0}) + \mathrm{TV}(\widehat{P}_s^{\mathcal{A}_D}, P_s^{w^1}) \geq 2\Delta_s.$$

Multiplying by μ_s and summing over $s \in H$ gives

$$\mathcal{R}_H(\mathcal{A}_D, w^0) + \mathcal{R}_H(\mathcal{A}_D, w^1) \geq 2 \sum_{s \in H} \mu_s \Delta_s.$$

Therefore,

$$\max_{j \in \{0,1\}} \mathcal{R}_H(\mathcal{A}_D, w^j) \geq \frac{1}{2} \left[\mathcal{R}_H(\mathcal{A}_D, w^0) + \mathcal{R}_H(\mathcal{A}_D, w^1) \right] \geq \sum_{s \in H} \mu_s \Delta_s.$$

Taking the infimum over all surface-level methods yields

$$\inf_{\mathcal{A}_D} \max_{j \in \{0,1\}} \mathcal{R}_H(\mathcal{A}_D, w^j) \geq \sum_{s \in H} \mu_s \Delta_s.$$

□

Corollary 12 (Mechanism Matching Breaks the Lower Bound)**.** *Let $H \subseteq \mathcal{S}_\star$ be a set of hidden candidate positions satisfying the assumptions of Lemma 11. Suppose further that mechanism-level evidence distinguishes the two worlds:*

$$\mathcal{O}_M^\infty(w^0) \neq \mathcal{O}_M^\infty(w^1),$$

and that mechanism matching retrieves analogies covering every $s \in H$ with transfer error

$$\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^{w^j}) \leq \alpha_s^{\mathrm{tr}}$$

for all $s \in H$ and $j \in \{0, 1\}$. Then

$$\max_j \mathcal{R}_H(\mathcal{A}_M, w^j) \leq \sum_{s \in H} \mu_s \alpha_s^{\mathrm{tr}}.$$

If

$$\sum_{s\in H}\mu_s\alpha_s^{\mathrm{tr}} < \sum_{s\in H}\mu_s\Delta_s,$$

then mechanism matching strictly dominates every surface method on hidden positions:

$$\inf_{\mathcal{A}_M}\max_j \mathcal{R}_H(\mathcal{A}_M, w^j) < \inf_{\mathcal{A}_D}\max_j \mathcal{R}_H(\mathcal{A}_D, w^j).$$

Proof. By the mechanism-transfer condition, for every $s \in H$ and every $j \in \{0,1\}$,

$$\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^{w^j}) \le \alpha_s^{\mathrm{tr}}.$$

Therefore,

$$\mathcal{R}_H(\mathcal{A}_M, w^j) = \sum_{s\in H}\mu_s \mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^{w^j}) \le \sum_{s\in H}\mu_s\alpha_s^{\mathrm{tr}}.$$

Taking the maximum over j preserves the same upper bound:

$$\max_j \mathcal{R}_H(\mathcal{A}_M, w^j) \le \sum_{s\in H}\mu_s\alpha_s^{\mathrm{tr}}.$$

By Lemma 11, every surface-level method satisfies

$$\inf_{\mathcal{A}_D}\max_j \mathcal{R}_H(\mathcal{A}_D, w^j) \ge \sum_{s\in H}\mu_s\Delta_s.$$

If

$$\sum_{s\in H}\mu_s\alpha_s^{\mathrm{tr}} < \sum_{s\in H}\mu_s\Delta_s,$$

then the mechanism-level upper bound lies strictly below the surface-level lower bound. Hence there exists a mechanism-level method whose worst-case hidden-position risk is strictly smaller than the worst-case risk of every surface-level method. □

Theorem 13 (Cross-Analogy Confirmation)**.** *Let $Z_s = \mathbb{1}[s \in \mathcal{S}_T^{\mathrm{true}}]$ and $X_{k,s} = \mathbb{1}[$analogy E_k confirms position $s]$. Let $\pi_s = P(Z_s = 1)$ and $X_{\mathcal{A},s} = \{X_{k,s} : E_k \in \mathcal{A}\}$.*

Assume $X_{k,s} \mid Z_s$ are conditionally independent with

$$q_{k,s} = P(X_{k,s} = 1 \mid Z_s = 1, E_k \in \mathcal{A}), \qquad p_{k,s} = P(X_{k,s} = 1 \mid Z_s = 0, E_k \in \mathcal{A}),$$

where $q_{k,s} > p_{k,s}$. The probabilities are understood conditional on the selection rule that produced $\mathcal{A}$.

Then:

$$\log\frac{P(Z_s = 1 \mid X_{\mathcal{A},s})}{P(Z_s = 0 \mid X_{\mathcal{A},s})} = \log\frac{\pi_s}{1-\pi_s} + \sum_{E_k\in\mathcal{A}}\left[X_{k,s}\log\frac{q_{k,s}}{p_{k,s}} + (1-X_{k,s})\log\frac{1-q_{k,s}}{1-p_{k,s}}\right].$$

The Bayes factor grows multiplicatively: each independent confirming analogy multiplies odds by $q_{k,s}/p_{k,s}$.

In the homogeneous positive-confirmation case, $q_{k,s} = q_s$, $p_{k,s} = p_s$, and $X_{k,s} = 1$ for all $E_k \in \mathcal{A}$, the number of confirmations needed to reach posterior confidence $1-\delta$ is

$$n_s^* = \left\lceil \frac{\log\frac{1-\delta}{\delta} - \log\frac{\pi_s}{1-\pi_s}}{\log\frac{q_s}{p_s}} \right\rceil.$$

Under the calibrated regime ($\pi_s = 0.5, q_s = 1, p_s = 0.2$), this gives $n_s^* = 2$ for $\delta = 0.05$.

For foresight, letting

$$\delta_s = P(Z_s = 0 \mid X_{\mathcal{A},s}),$$

the posterior expected risk at position s satisfies

$$\mathbb{E}[\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^T) \mid X_{\mathcal{A},s}] \le (1-\delta_s)\alpha_s^{\mathrm{tr}} + \delta_s.$$

Proof. Let

$$X_{\mathcal{A},s} = \{X_{k,s} : E_k \in \mathcal{A}\}.$$

For a fixed realization $x_{\mathcal{A},s}$, define the likelihood under the structural hypothesis $Z_s = 1$:

$$L_1(x_{\mathcal{A},s}) = P(X_{\mathcal{A},s} = x_{\mathcal{A},s} \mid Z_s = 1).$$

By conditional independence,

$$L_1(x_{\mathcal{A},s}) = \prod_{E_k \in \mathcal{A}} q_{k,s}^{x_{k,s}} (1 - q_{k,s})^{1 - x_{k,s}}.$$

Similarly, under the coincidence hypothesis $Z_s = 0$,

$$L_0(x_{\mathcal{A},s}) = P(X_{\mathcal{A},s} = x_{\mathcal{A},s} \mid Z_s = 0) = \prod_{E_k \in \mathcal{A}} p_{k,s}^{x_{k,s}} (1 - p_{k,s})^{1 - x_{k,s}}.$$

By Bayes' theorem,

$$P(Z_s = 1 \mid X_{\mathcal{A},s}) = \frac{\pi_s L_1(X_{\mathcal{A},s})}{\pi_s L_1(X_{\mathcal{A},s}) + (1 - \pi_s) L_0(X_{\mathcal{A},s})}.$$

Likewise,

$$P(Z_s = 0 \mid X_{\mathcal{A},s}) = \frac{(1 - \pi_s) L_0(X_{\mathcal{A},s})}{\pi_s L_1(X_{\mathcal{A},s}) + (1 - \pi_s) L_0(X_{\mathcal{A},s})}.$$

Therefore the posterior odds are

$$\frac{P(Z_s = 1 \mid X_{\mathcal{A},s})}{P(Z_s = 0 \mid X_{\mathcal{A},s})} = \frac{\pi_s}{1 - \pi_s} \cdot \frac{L_1(X_{\mathcal{A},s})}{L_0(X_{\mathcal{A},s})}.$$

The likelihood ratio is

$$\frac{L_1(X_{\mathcal{A},s})}{L_0(X_{\mathcal{A},s})} = \prod_{E_k \in \mathcal{A}} \left(\frac{q_{k,s}}{p_{k,s}}\right)^{X_{k,s}} \left(\frac{1 - q_{k,s}}{1 - p_{k,s}}\right)^{1 - X_{k,s}}.$$

Taking logarithms gives

$$\log \frac{P(Z_s = 1 \mid X_{\mathcal{A},s})}{P(Z_s = 0 \mid X_{\mathcal{A},s})} = \log \frac{\pi_s}{1 - \pi_s} + \sum_{E_k \in \mathcal{A}} \left[X_{k,s} \log \frac{q_{k,s}}{p_{k,s}} + (1 - X_{k,s}) \log \frac{1 - q_{k,s}}{1 - p_{k,s}} \right].$$

If $X_{k,s} = 1$ for every $E_k \in \mathcal{A}$, then the Bayes factor reduces to

$$BF_s(\mathcal{A}) = \prod_{E_k \in \mathcal{A}} \frac{q_{k,s}}{p_{k,s}}.$$

Thus every independent confirming analogy multiplies the posterior odds by $q_{k,s}/p_{k,s}$.

In the homogeneous case $q_{k,s} = q_s$ and $p_{k,s} = p_s$, with K positive confirmations,

$$BF_s(K) = \left(\frac{q_s}{p_s}\right)^K.$$

The posterior odds are

$$\frac{P(Z_s = 1 \mid K \text{ positives})}{P(Z_s = 0 \mid K \text{ positives})} = \frac{\pi_s}{1 - \pi_s} \left(\frac{q_s}{p_s}\right)^K.$$

To achieve posterior confidence at least $1 - \delta$, we require

$$\frac{\pi_s}{1 - \pi_s} \left(\frac{q_s}{p_s}\right)^K \geq \frac{1 - \delta}{\delta}.$$

Solving for K yields

$$K \geq \frac{\log \frac{1-\delta}{\delta} - \log \frac{\pi_s}{1-\pi_s}}{\log \frac{q_s}{p_s}}.$$

Hence the minimum integer number of confirmations is

$$n_s^* = \left\lceil \frac{\log \frac{1-\delta}{\delta} - \log \frac{\pi_s}{1-\pi_s}}{\log \frac{q_s}{p_s}} \right\rceil.$$

For the calibrated regime

$$\pi_s = 0.5, \qquad q_s = 1, \qquad p_s = 0.2, \qquad \delta = 0.05,$$

we have prior odds equal to 1, and the required posterior odds are

$$\frac{1-\delta}{\delta} = \frac{0.95}{0.05} = 19.$$

Since

$$\frac{q_s}{p_s} = 5,$$

one confirmation gives odds $5 < 19$, while two confirmations give odds $25 > 19$. Therefore

$$n_s^* = 2.$$

For the foresight-risk bound, define

$$\delta_s = P(Z_s = 0 \mid X_{\mathcal{A},s}).$$

With posterior probability $1 - \delta_s$, position s is truly structural. Under this event, the mechanism-transfer assumption gives

$$\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^T) \leq \alpha_s^{\mathrm{tr}}.$$

With posterior probability δ_s, the inferred position is coincidental or otherwise not truly structural. In that case we use the trivial upper bound

$$\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^T) \leq 1.$$

Therefore,

$$\mathbb{E}\left[\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^T) \mid X_{\mathcal{A},s}\right] \leq (1-\delta_s)\alpha_s^{\mathrm{tr}} + \delta_s \cdot 1.$$

Thus,

$$\mathbb{E}\left[\mathrm{TV}(\widehat{P}_s^{\mathcal{A}_M}, P_s^T) \mid X_{\mathcal{A},s}\right] \leq (1-\delta_s)\alpha_s^{\mathrm{tr}} + \delta_s.$$

This separates the two sources of foresight error: identification error δ_s and transfer error α_s^{tr}. □

D. More Details of `ADR-bench`

We list the rubric prompts used in `ADR-bench` as below.

Call 2 – Per-Analogy Scoring (AVS) & Cross-Analogy Reasoning (CARS) prompt

<task>
You are evaluating the quality of analogical reasoning in a report. This covers BOTH individual analogy quality AND cross-analogy synthesis.

TARGET EVENT: {event_name} (as of {cutoff_date})
ANALOGIES THE SYSTEM FOUND: {extracted_analogies}
REFERENCE SET (expert-curated analogies from academic literature): {reference_set}
</task>
<part-A: per-analogy scoring (AVS)>
For each analogy in the system's list, evaluate:

1. factual_existence: "PASS" if real historical event, "FAIL" if hallucinated.

2. structural_relevance (0–4) — how complete is the causal mechanism mapping?
0 = No structural relevance; pure surface match (same country/topic only).
1 = Shares one superficial feature ("both involve banks").
2 = Identifies 1–2 genuine shared mechanisms but without causal chain depth.
3 = Identifies 3+ shared mechanisms with causal chain articulation (trigger → mediator → outcome).
4 = Complete structural mapping: preconditions, causal chains, mechanisms, and outcomes all explicitly aligned between source and target. Must name specific institutional actors or variables in BOTH events.

Calibration — Score 4: "Trust companies (1907) operated outside the National Banking System with ∼5% reserves vs 25% for national banks. Today's SIVs/conduits similarly operate outside FDIC coverage and bank capital requirements. Both were funded by demandable/rollable short-term liabilities (deposits then, ABCP now) against illiquid long-term assets. The structural role is identical: maturity-transforming intermediary outside the safety net."
Calibration — Score 2: "Like the S&L crisis, this involves real estate losses leading to financial institution failures."

3. is_in_reference_set: "YES" if matches a reference event, "PARTIAL" if sibling/related, "NO" otherwise.

4. novelty: "NOT_NOVEL" if in reference set; "NOVEL_VALID" if real event with structural_relevance ≥ 2 not in reference set; "NOVEL_INVALID" otherwise.

5. difference_awareness (0–2):
0 = No limitations discussed.
1 = Mentions differences superficially ("the scale is different").
2 = Identifies WHERE the analogy breaks down and WHY it matters for the analysis.
</part-A>
<part-B: cross-analogy reasoning (CARS)>
Evaluate how the report uses multiple analogies TOGETHER. This is the critical test — most reports list analogies pairwise but never synthesize across them.

B1. Invariant Extraction Quality (0–4). Sub-classify each cross-analogy invariant the report identifies:
L3-D (descriptive): "All three events share feature X" — shared attribute without specifying STRUCTURAL ROLE (trigger? amplifier? enabler? mediator?). Score weight: 0.4 per invariant.
L3-S (structural): X occupies the SAME STRUCTURAL ROLE in all events' causal chains, even when instantiated by different domain-specific variables. Score weight: 1.0 per invariant.

Calibration — L3-S: "Across 1907 (Heinze director interlocks), Japan 1990 (keiretsu cross-shareholdings), LTCM 1998 (OTC counterparty opacity), an opaque interconnection network serves as the AMPLIFIER that transforms localized shock into systemic crisis."
Calibration — L3-D: "In every case — 1907, 1929, 2008 — institutions built up huge leverage on illiquid assets."

Score	Criteria
0	No cross-analogy comparison attempted
1	Notes shared features but no systematic analysis
2	Identifies 1–2 invariants, mainly L3-D
3	Identifies 3+ invariants, at least 1 is L3-S
4	Systematic extraction + explicit universal vs domain-specific distinction + $\geq$2 L3-S invariants

B2. Hidden Factor Inference (0–3). Does the report infer SPECIFIC hidden factors in the TARGET EVENT based on patterns across multiple analogies?

Score	Criteria
0	No inference of unobserved factors
1	Generic statement ("there may be hidden risks") — L4-mention
2	Names specific hidden factor + $\geq$1 analogy justification, but without completing cross-analogy derivation — L4-partial
3	Complete derivation: (a) names structural role recurring across analogies, (b) cites $\geq$2 analogies with specific instantiations, (c) infers specific hidden factor in target, (d) explains why it hasn't been identified yet — L4-full

B3. Analogy Differentiation (0–2).

Score	Criteria
0	Treats all analogies as interchangeable supporting examples
1	Notes different analogies highlight different aspects
2	Explicitly maps WHICH insight comes from WHICH analogy and WHY certain analogies are more informative for specific aspects

</part-B>
<output>
Output exactly ONE JSON object:

```
{
 "per_analogy": [{"event_name": "...", "factual_existence": "PASS|FAIL",
  "structural_relevance": 0, "is_in_reference_set": "YES|PARTIAL|NO",
  "novelty": "NOT_NOVEL|NOVEL_VALID|NOVEL_INVALID",
  "difference_awareness": 0}],
 "cross_analogy_reasoning": {
  "B1_invariant_extraction": {"score": 0, "L3_S_count": 0, "L3_D_count": 0,
"justification": "..."},
  "B2_hidden_factor_inference": {"score": 0, "L4_claims": [...], "justification":
"..."},
  "B3_analogy_differentiation": {"score": 0, "justification": "..."},
  "CARS": (B1+B2+B3)/9 },
 "aggregate": { "structural_precision": 0.0, "reference_recall": 0.0,
  "cross_domain_rate": 0.0, "novel_valid_count": 0,
  "mean_structural_relevance": 0.0,
  "mean_AVS": mean of (struct_rel + diff_aware + novelty_bonus)/10,
  "ARQ": mean_AVS×0.6 + CARS×0.4 } }
```

</output>

Call 3 – Atomic Claim Decomposition (L1–L4) & Hidden-Factor Detection prompt

<task>
Decompose this analysis report into atomic analytical claims. Each claim should be a single assertion that can be independently evaluated. For each claim, classify its level.
</task>
<rubric>
L1 (Surface Fact): A factual observation about the current situation, available from basic news/data. No historical reasoning involved. Also includes name-dropping a historical event WITHOUT mechanism mapping.
Example: "Subprime mortgage delinquencies have been rising since late 2006."
Example (name-drop): "This is reminiscent of the 2008 financial crisis." [No mechanism mapping → L1]

L2 (Single-Analogy Insight): An insight that explicitly draws on ONE specific historical event and maps a CONCRETE MECHANISM from that event to the current situation. Must name what structural element is shared and how it operates in both contexts.
Example: "Like the 1907 trust companies, today's SIVs operate outside the lender-of-last-resort perimeter, making them vulnerable to wholesale funding runs." [Specific mechanism: LoLR perimeter gap → run vulnerability]

L3-D (Cross-Analogy Descriptive Pattern): Notes that multiple ($\geq$2) historical events share a common FEATURE, but without specifying the STRUCTURAL ROLE that feature plays.
Example: "In every major crisis — 1907, 1929, 2008 — excess leverage was present." [Shared feature, but is leverage a trigger, amplifier, or enabler?]

L3-S (Cross-Analogy Structural Pattern): Identifies that multiple ($\geq$2) historical events share a mechanism occupying the SAME STRUCTURAL ROLE in the causal chain, even when instantiated by different domain-specific variables. Must explicitly name or clearly imply the structural role.
Example: "Across 1907 (Heinze director interlocks), Japan 1990 (keiretsu cross-shareholdings), and LTCM 1998 (OTC counterparty opacity), an opaque interconnection network serves as the AMPLIFIER that transforms a localized shock into systemic crisis."

L4 (Hidden Factor Inference): An inference about a SPECIFIC, CURRENTLY UNRECOGNIZED factor in the current situation, justified by cross-analogy evidence from $\geq$2 historical events. ALL of the following must be present:
(a) Names a SPECIFIC factor or mechanism (not "there may be hidden risks")
(b) Explains WHY this factor is likely present, citing $\geq$2 analogies

(c) The factor must be something not yet widely discussed in pre-cutoff reporting
(d) Must be a FACTOR or MECHANISM identification, NOT a probabilistic forecast or scenario prediction

Valid L4: "Based on the recurring pattern of opaque amplification networks in 1907/1998, an analogous hidden interconnection likely exists — possibly CDS counterparty exposure concentrated in a few dealers, or repo market concentration in clearing banks."
NOT L4 (probabilistic forecast → L1): "There is a 70% probability of recession."
NOT L4 (scenario → L1): "Based on historical patterns, a 30–40% market correction is likely."
NOT L4 (too vague → L1): "There may be hidden systemic risks."
NOT L4 (known factor → L1): "The ABCP market freeze is causing liquidity problems."
</rubric>
<critical-rules>
1. Probabilistic forecasts and scenario predictions are NEVER L4, regardless of historical references. L4 identifies a HIDDEN FACTOR, not an OUTCOME.
2. Name-dropping a historical event without mechanism mapping is L1, not L2.
3. Noting a shared feature across events without specifying its structural role is L3-D, not L3-S.
4. A claim can only be L3-S or L4 if it references MULTIPLE (≥2) historical events.
5. If unsure between L3-D and L3-S, check: does the claim name or clearly imply a structural role (trigger, amplifier, mediator, enabler, outcome)? If yes → L3-S. If no → L3-D.
</critical-rules>
<input>
REPORT: {report}
HIDDEN FACTORS FROM REFERENCE SET (use to check if any claims — at ANY level — match known hidden factors): {hidden_factors}
</input>
<output>
```
{ "claims": [{
 "claim": "exact text from report (may be condensed)",
 "level": "L1 | L2 | L3-D | L3-S | L4",
 "historical_events_referenced": ["event1", "event2"],
 "mechanism_identified": "..." or null,
 "structural_role": "trigger|amplifier|mediator|enabler|outcome" or null,
 "hidden_factor_hit": "name from reference set" or null }],
"aggregate": { "L1": 0, "L2": 0, "L3_D": 0, "L3_S": 0, "L4": 0,
 "total_claims": 0,
 "HF_at_any": "count of distinct HFs hit at ANY level",
 "HF_at_L4": "count of distinct HFs hit at L4 ONLY",
 "HF_total": "total HFs in reference set" } }
```
</output>

Call 4 – Reasoning Analysis Score (RAS) prompt

<task>
You are evaluating the analytical quality of a report about a historical situation. Score on six dimensions.

EVENT: {event_name}; TEMPORAL CUTOFF: {cutoff_date}
The analyst was asked to write a comprehensive analysis as of the cutoff date.
REPORT: {report}

Score each dimension according to the rubric. Be STRICT and CALIBRATED.
</task>
<calibration-guidance>
- Scores of 3–4 (on 0–4 scales) or 3 (on 0–3 scales) should be RARE. They require genuinely exceptional analysis that goes beyond competent reporting.
- A well-written report by a frontier LLM that describes mechanisms competently but without historical grounding or structural novelty should typically score 2–3 on D1 and 1–2 on D4.
- The TOP score on each dimension requires something a generic deep-research report would NOT naturally produce — typically historical grounding, structural abstraction, or calibrated uncertainty from precedent.
</calibration-guidance>
<D1: causal depth (0–4, weight ×2)>
0: Purely descriptive; no causal claims.
1: Identifies causes but only immediate/proximate.

2: Identifies causal chains with 2+ steps (A→B→C); names some structural features.
3: Multi-level causal analysis: distinguishes proximate causes from structural causes and enabling conditions; identifies feedback loops and amplification mechanisms.
4: Everything in 3, PLUS: explicitly names the structural role each factor plays (trigger, amplifier, mediator, enabler); grounds the causal framework in historical precedent showing WHY these specific mechanisms should be expected; distinguishes this event's causal structure from superficially similar but structurally different events.

Calibration — Score 3: "The crisis operates through a self-reinforcing funding liquidity spiral: asset sales depress prices, triggering margin calls, forcing more sales." [Identifies feedback loop — but no historical grounding or structural role labeling]
Calibration — Score 4: "Three feedback loops operate simultaneously. Loop 1 (mark-to-market → haircut): ABX prices fall → dealers mark books → repo lenders widen haircuts → forced sales push ABX lower. This is structurally identical to the 1907 trust-company run mechanism, where opacity about which trusts held Heinze-linked assets triggered indiscriminate withdrawal. The TRIGGER is subprime losses; the AMPLIFIER is opacity-driven wholesale funding withdrawal; the MEDIATOR is the ABCP/repo conduit system."
</D1>
<D2: mechanism specificity (0–3, weight ×2)>
0: No specific mechanisms named.
1: Names mechanisms generically ("contagion", "panic", "leverage", "deleveraging").
2: Specifies concrete transmission channels with institutional detail.
3: Specifies channels + quantitative data + named institutional actors.

Calibration — Score 2: "Banks are tightening lending to each other, as shown by widening LIBOR-OIS spreads."
Calibration — Score 3: "The effective federal funds rate printed at 5.41% on Aug 9, 16bp above the 5.25% target — the largest deviation in over a year. The ECB allotted €94.8B to 49 banks at 4.00% full allotment."
</D2>
<D3: scope and coverage (0–3, weight ×2)>
0: Covers only one dimension of the event.
1: Covers multiple dimensions but misses major ones.
2: Comprehensive coverage of known risk dimensions (housing, securitization, shadow banking, interbank, policy response).
3: Comprehensive + identifies under-discussed or emerging dimensions a standard analysis would miss (e.g., quant fund deleveraging as second-order effect; European bank dollar-funding dependence; MMF structural vulnerability; commercial real estate as next domino).
</D3>
<D4: nuance and uncertainty (0–3, weight ×1)>
0: Entirely confident assertions with no hedging.
1: Some uncertainty language but generic.
2: Specific uncertainty about specific claims with reasoning.
3: Distinguishes high- from low-confidence claims; CALIBRATES uncertainty using specific historical precedent; identifies what specific information would change the assessment.

Calibration — Score 2: "The outcome is uncertain and depends on policy responses and the depth of housing correction."
Calibration — Score 3: "If LIBOR-OIS stays above 100bp, escalate to 1907 trajectory and expect multi-month dislocation; if ABCP issuance resumes within a month, the 1998 template applies — sharp dislocation, swift resolution. The 1907/1998 distinction hinges on whether the problem is solvency (1907) or pure liquidity (1998)."
</D4>
<D5: internal consistency (0–2, weight ×1)>
0: Contains contradictions or logically incompatible claims.
1: No contradictions but claims are loosely connected; analysis reads as a list of observations.
2: Claims form a coherent, mutually reinforcing analytical framework with a clear through-line.
</D5>
<D6: actionability (0–2, weight ×1)>
0: Analysis is purely retrospective/academic with no forward-looking implications.
1: Identifies what to monitor but not specific thresholds or decision triggers.
2: Provides specific indicators with thresholds, scenario-conditional recommendations, decision-relevant framings (e.g., monitoring table with current value → escalation threshold → normalization threshold for LIBOR-OIS, ABCP outstanding, ABX.HE.AAA, etc.).
</D6>
<output>
```
{ "D1_causal_depth": {"score": 0, "justification": "..."},
 "D2_mechanism_specificity": {"score": 0, "justification": "..."},
 "D3_scope": {"score": 0, "justification": "..."},
```

```
 "D4_nuance": {"score": 0, "justification": "..."},
 "D5_consistency": {"score": 0, "justification": "..."},
 "D6_actionability": {"score": 0, "justification": "..."},
 "RAS": (D1×2 + D2×2 + D3×2 + D4 + D5 + D6) / 27 }
</output>
```

Call 5 – Foresight Quality Score (FQS) prompt

<task>
You are evaluating the foresight quality of an analysis report written about a historical situation AS OF a specific cutoff date. You know what actually happened after the cutoff — use this knowledge to evaluate the report's forward-looking claims.

EVENT: {event_name}; TEMPORAL CUTOFF: {cutoff_date}
The analyst was asked to provide forward-looking assessment as of the cutoff date.
REPORT: {report}
KNOWN HIDDEN FACTORS (factors not widely recognized at cutoff but proved critical): {hidden_factors}
</task>
<step-1: extract>
Extract every claim that makes a prediction, projects a scenario, or assesses future risk. Include conditional statements ("if X, then Y"), probability assessments, and scenario descriptions. Be exhaustive.
</step-1>
<step-2: score each claim>
A. OUTCOME (verified against what actually happened):
+1.0 Confirmed: development occurred substantially as predicted.
+0.5 Partially confirmed: directionally correct but details differ materially.
+0.0 Unconfirmed: did not occur (but not contradicted).
−0.3 Contradicted: opposite occurred or fundamentally wrong mechanism.

B. GROUNDING (0–3) — BE STRICT. This is the key discriminating dimension.
0: No historical analogy or evidence cited. Pure assertion or extrapolation. Includes "historically, markets recover" or "past crises suggest..." without naming a SPECIFIC event.
1: Names a specific historical event but provides NO mechanism mapping ("reminiscent of 1998"). The event is decoration, not analytical input.
2: Names a specific event AND aligns ONE structural element (trigger OR mechanism OR outcome — but not all). Example: "Following the 1907 template, trust-company-style runs on shadow banks are likely."
3: Names a specific event AND aligns $\geq$2 structural roles AND names the corresponding variable in the CURRENT situation. Example: "Based on 1907 (trust companies outside NYCH = TRIGGER: opacity-driven run), the current TRIGGER is ABCP conduit opacity; the AMPLIFIER is the same (maturity mismatch + wholesale funding); the likely OUTCOME, per the 1907 template, is sequential institutional failures until a coordinator (Morgan then, Fed now) provides selective support."

Calibration note: Generic references to "history shows" or "past crises" with no specific event = Grounding 0. Naming "1998" without saying WHAT about 1998 applies = Grounding 1. Most foresight claims by non-synthesis agents will be Grounding 0 or 1.

C. SPECIFICITY (1–3):
1: Vague/directional ("financial system will face significant stress").
2: Names a causal pathway ("credit contraction via bank deleveraging will slow GDP growth").
3: Names concrete entities, quantities, or timing with mechanism ("investment banks with subprime MBS exposure $>10\times$ capital face solvency risk within 6 months; the Bear Stearns funds are the template").

D. DERIVABILITY (0–3):
0: Impossible without post-cutoff knowledge $\rightarrow$ temporal violation.
1: Unlikely derivable from pre-cutoff evidence alone.
2: Plausible stretch from pre-cutoff evidence + analogical reasoning.
3: Clearly derivable from pre-cutoff evidence + analogies.

E. TEMPORAL COMPLIANCE:
CLEAN: Based on pre-cutoff facts or analogical reasoning (citing full history of analogy events is allowed).
SOFT_VIOLATION: References post-cutoff trends about the TARGET event without naming specific post-cutoff events.
HARD_VIOLATION: Explicitly references specific post-cutoff events, data, or outcomes about the TARGET event.

F. HIDDEN FACTOR HIT: Does this claim identify or predict any of the known hidden factors listed above? A hit requires the claim to identify the FACTOR, not just predict an outcome that happens to relate to it.

</step-2>
<step-3: per-claim FQS>
$w = 1.0$ if CLEAN, 0.5 if SOFT_VIOLATION, 0.0 if HARD_VIOLATION.
$\text{RQ} = (\text{Grounding} + \text{Specificity} + \text{Derivability})/9$
$\text{FQS} = w \times (\text{Outcome} \times 0.5 + \text{RQ} \times 0.5)$
</step-3>
<step-4: aggregate & output>

```
{ "foresight_claims": [{
  "claim": "...", "outcome": {"rating": "confirmed|partial|unconfirmed|contradicted",
    "value": 0.0, "what_actually_happened": "..."},
  "grounding": {"score": 0, "analogy_cited": "..." or null,
    "structural_roles_mapped": ["trigger","mechanism","outcome"] or [],
    "justification": "..."},
  "specificity": {"score": 0, "justification": "..."},
  "derivability": {"score": 0, "justification": "..."},
  "temporal_compliance": "CLEAN|SOFT_VIOLATION|HARD_VIOLATION",
  "hidden_factor_hit": "..." or null, "RQ": 0.0, "FQS": 0.0 }],
"aggregate": { "total_foresight_claims": 0, "FQS_mean": 0.0,
  "FQS_depth": "mean FQS over claims with Grounding ≥ 2 AND Specificity ≥ 2",
  "FQS_depth_n": "count meeting that threshold",
  "temporal_compliance_rate": 0.0, "confirmed_rate": 0.0,
  "grounded_rate": "fraction with Grounding ≥ 2",
  "hidden_factor_hits": 0, "hidden_factor_total": 0 } }
```

</step-4>

E. More Details of CANA

E.1. CANA for historical analogy generation

The algorithmic description of CANA used for

Algorithm 1 The CANA Framework

Required: Target event E_T; candidate analogy universe $\mathcal{U}$; LLM $\mathcal{M}$; representation extractor $\textsc{Repr}(\cdot)$ that produces $(D(E), M(E))$; pattern abstractor $\textsc{Abs}(\cdot)$ that produces the abstract causal pattern $\mathfrak{P}$ and its structural positions $\mathcal{S}_{\mathfrak{P}}$; candidate generator $\textsc{Gen}(\cdot)$; self-analogy filter $\textsc{Filter}(\cdot,\cdot)$; per-position alignment scorer $\textsc{Align}(\cdot,\cdot)$ that produces $\{X_{k,s}\}_{s\in\mathcal{S}_{\mathfrak{P}}}$ for analogy E_k; aggregate decider $\textsc{Decide}(\cdot)$; max reflection rounds R; candidates per round K; carryover size M;
Extracting the target representations $(D(E_T), M(E_T)) \leftarrow \textsc{Repr}(\mathcal{M}, E_T)$;
Abstracting the target pattern $(\mathfrak{P}, \mathcal{S}_{\mathfrak{P}}) \leftarrow \textsc{Abs}(\mathcal{M}, D(E_T), M(E_T))$, where $\mathcal{S}_{\mathfrak{P}}$ enumerates the structural positions to be confirmed by analogies;
Sampling an initial candidate analogy set $\mathcal{A}^{(0)} \subset \mathcal{U}$ via $\mathcal{A}^{(0)} \leftarrow \textsc{Gen}(\mathcal{M}, \mathfrak{P}, K)$;
Initializing the tried-set $\mathcal{T} \leftarrow \mathcal{A}^{(0)}$, the carryover pool $\mathcal{K} \leftarrow \emptyset$, and the latest decision $\mathcal{D}^* \leftarrow$ `null`;
for round $t \in [0, 1, \ldots, R]$ **do**
  Initializing the per-round evaluation set $\mathcal{E}^{(t)} \leftarrow \emptyset$;
  for each candidate analogy $E_k \in \mathcal{A}^{(t)}$ **do**
    if $\textsc{Filter}(\mathcal{M}, E_T, E_k)$ flags E_k as alias, sub-instance, or proper subset of E_T, i.e., E_k collapses to E_T rather than residing in $\mathcal{U} \setminus \{E_T\}$ in the strict sense **then**
      Discarding E_k;
    else
      Extracting $(D(E_k), M(E_k)) \leftarrow \textsc{Repr}(\mathcal{M}, E_k)$ via Wikipedia-grounded retrieval followed by mechanism extraction;
      Producing the position-level alignment evidence $\{X_{k,s}\}_{s\in\mathcal{S}_{\mathfrak{P}}} \leftarrow \textsc{Align}(\mathcal{M}, (D(E_T), M(E_T)), (D(E_k), M(E_k)))$, where each $X_{k,s}$ records both the confirming match and the residual gap of analogy E_k at position s;
      Updating $\mathcal{E}^{(t)} \leftarrow \mathcal{E}^{(t)} \cup \{(E_k, \{X_{k,s}\}_s)\}$;
    end if
  end for
  Merging the carryover from the previous round: $\mathcal{E}^{(t)} \leftarrow \mathcal{E}^{(t)} \cup \mathcal{K}$;
  Computing the aggregate decision $\mathcal{D}^{(t)} \leftarrow \textsc{Decide}(\mathcal{M}, \mathfrak{P}, \mathcal{E}^{(t)})$, which (i) prunes candidates violating five hard rejection rules (alias, sub-event, super-set, generic descriptor, scope-class mismatch), (ii) ranks the eligible analogies by the fidelity of the induced structural mapping ϕ_k from E_k to E_T, equivalently by the support that $\{X_{k,s}\}_s$ provides for $\widehat{P}_s^{\mathcal{A}}$ across $s \in \mathcal{S}_{\mathfrak{P}}$ under mechanism specificity and scale alignment, and (iii) returns $\{E^{(t)*}, \text{top-}M, \text{is_sufficient}, \text{common_gap}, \text{next_focus}\}$;
  Updating $\mathcal{D}^* \leftarrow \mathcal{D}^{(t)}$;
  if $\mathcal{D}^{(t)}$.is_sufficient holds, indicating that the target-side overlap Ω_{pos} is adequately covered by the chosen analogy **then**
    break;
  end if
  Updating the carryover pool $\mathcal{K} \leftarrow \mathcal{D}^{(t)}$.top-$M$;
  Generating the next-round candidate analogies $\mathcal{A}^{(t+1)} \leftarrow \textsc{Gen}(\mathcal{M}, \mathfrak{P}, \mathcal{D}^{(t)}.\text{common_gap}, \mathcal{D}^{(t)}.\text{next_focus}, \mathcal{T})$, conditioned on the under-confirmed positions in $\mathcal{S}_{\mathfrak{P}}$ and excluding any $E_k \in \mathcal{T}$;
  Removing duplicates: $\mathcal{A}^{(t+1)} \leftarrow \mathcal{A}^{(t+1)} \setminus \mathcal{T}$;
  if $\mathcal{A}^{(t+1)} = \emptyset$ **then**
    break;
  end if
  Updating the tried-set $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{A}^{(t+1)}$;
end for
return the selected analogy $E^* \leftarrow \mathcal{D}^*.E^{(t)*}$ together with its induced structural mapping $\phi^* : \mathcal{S}_{\mathfrak{P}} \to E^*$;

We present the prompts we used in CANA:

Candidate Generation prompt

<task>
You are a robot for finding structurally analogous historical events based on causal analysis. Historical Analogy means identifying past events that share similar causal structures with a contemporary event, not merely similar topics or surface features.

Given the causal profile of an input event (including its structural pattern, mechanism, and causal chain), output 5 historical events that share similar causal structures. Candidates may come from entirely different domains (e.g., a financial crisis may be structurally analogous to an epidemic) as long as the underlying causal mechanisms are similar. If there is any reflection, please modify the recommended events based on the reflection.
</task>
<format>
IMPORTANT: Output ONLY a JSON list of 5 short event names, like `[``event1'',``event2'',``event3'',``event4'',``event5''`
No explanations, no descriptions, no numbering. Each name should be a concise historical event name (under 50 characters).
</format>
<example>
Input Event:
2008 Financial Crisis
Structural Pattern: systemic_risk_cascade
Mechanism: Interconnected financial instruments amplified localized failures into global systemic collapse
Causal Chain: Housing bubble + lax regulation → subprime mortgage defaults → CDO/CDS losses cascade → bank insolvency → credit freeze → global recession
Output: ["1997 Asian financial crisis", "South Sea Bubble", "Long-Term Capital Management collapse", "2010 European sovereign debt crisis", "1873 Vienna Stock Exchange crash"]
</example>
<history>
{chat_history}
</history>
<input>
{input_type}: {input}
Output:
</input>

Self-Analogy / Alias / Sub-instance Filter prompt

<task>
You will judge whether a candidate event should be FILTERED OUT as a self-analogy when proposing it as an analog for an input event.

Input event: {input_name}
Candidate event: {candidate_name}
</task>
<rules>
Return YES (filter out) ONLY if one of these three holds:

1. ALIAS — candidate is a different NAME for the SAME specific historical event.
YES patterns (templated; angle brackets denote slots, not specific events):
– ⟨event under codename⟩ ≡ ⟨same event under descriptive popular name⟩
– ⟨event under regional/local name⟩ ≡ ⟨same event under translated or global name⟩
– ⟨event under full institutional suffix⟩ ≡ ⟨same event under truncated suffix or root noun⟩
– ⟨war named by calendar A date⟩ ≡ ⟨same war named by calendar B date⟩
– ⟨crisis named by location⟩ ≡ ⟨same crisis named by month/year⟩

2. SUB-INSTANCE of GENERIC input — input names a generic CATEGORY, candidate is one specific instance.
YES patterns:
– input "⟨generic ceremony class⟩" + candidate "⟨a regional or cultural variant of that ceremony⟩"
– input "⟨class of recurring institutional events⟩" + candidate "⟨one dated instance of that class⟩"
– input "⟨class of accidents/disasters⟩" + candidate "⟨one named incident of that kind⟩"

3. PROPER SUBSET — input is a CONTAINER event (a multi-year war, movement, era, crisis, or campaign that comprises many sub-events) and the candidate is one named sub-event WITHIN it.
YES patterns (true proper subsets):

– input "⟨multi-year war⟩" + candidate "⟨one named battle within that war⟩"
– input "⟨decade-long geopolitical confrontation⟩" + candidate "⟨one named crisis episode within that confrontation⟩"
– input "⟨social movement spanning years⟩" + candidate "⟨one named march/protest/campaign within that movement⟩"
– input "⟨multi-stage operation⟩" + candidate "⟨one named stage of that operation⟩"

COUNTER-EXAMPLES — do NOT filter these (sibling, not subset):
– input "⟨conference A in a series⟩" + candidate "⟨conference B in the same series⟩" → NO
– input "⟨war A⟩" + candidate "⟨separate war between different parties, similar mechanism⟩" → NO
– input "⟨war A⟩" + candidate "⟨war B as a sibling in a numbered series⟩" → NO

If unsure between SAME-DOMAIN-SIBLING (NO) and PROPER-SUBSET (YES), ask: *"Is the candidate event part of the input event's timeline / agenda / formal scope?"* — if no, it is a sibling, not a subset.

Return NO for everything else (genuinely different events sharing vocabulary; different events in same domain; cross-domain analogs).
</rules>
<reasoning>
Reason step by step BEFORE giving the final answer:
1. Identify the specific historical event each name refers to. Are they the same real-world event under different names?
2. If different, check: is one a SUB-INSTANCE of a generic category named by the other?
3. If different, check: is one a PROPER SUBSET (named sub-event within a container event)?
4. State your final decision.
</reasoning>
<output>
Reasoning: ⟨2–3 sentences⟩
Answer: YES or NO
</output>

Per-Candidate 5-Dimension Match/Gap Evaluation prompt

<task>
You will compare an input event against a candidate analogy along 5 causal dimensions. For each dimension, identify both what MATCHES and what does NOT match.
</task>
<input-event>
Input Event: {input_name}
Preconditions: {input_precond}
Causal Chain: {input_chain}
Mechanism: {input_mechanism}
Outcome & Impact: {input_outcome}
Structural Pattern: {input_pattern}
</input-event>
<candidate-event>
Candidate Event: {cand_name}
Preconditions: {cand_precond}
Causal Chain: {cand_chain}
Mechanism: {cand_mechanism}
Outcome & Impact: {cand_outcome}
Structural Pattern: {cand_pattern}
</candidate-event>
<reasoning>
REASONING-FIRST PROCESS — for EACH dimension, work through these steps:
Step 1: Read input's value for this dimension and candidate's value.
Step 2: Identify what specifically ALIGNS — name the shared mechanism / role / dynamic. Be concrete.
Step 3: Identify what specifically does NOT align — name the divergence.
Step 4: Fill the JSON match/gap fields based on Step 2/3 reasoning.

Rules:
- If a dimension has no meaningful match, set MATCH to "" (empty string).
- If a dimension matches with no important gap, set GAP to "" (empty string).
- Avoid hedging like "somewhat" or "loosely" — be specific.
- Do not output the reasoning steps; only the final JSON.

</reasoning>
<output>
Output ONLY valid JSON, no other text:

```
{
 "preconditions":   {"match": "...", "gap": "..."},
 "causal_chain":    {"match": "...", "gap": "..."},
 "mechanism":       {"match": "...", "gap": "..."},
 "outcome_impact":  {"match": "...", "gap": "..."},
 "structural_pattern": {"match": "...", "gap": "..."}
}
```

</output>

Aggregate Decision prompt (REJECT rules + ranking)

<task>
You have evaluated {n_cands} candidate analogies for the input event below. The per-candidate match/gap summaries are provided.
</task>
<input-event>
Input Event: {input_name}
Preconditions: {input_precond}
Causal Chain: {input_chain}
Mechanism: {input_mechanism}
Outcome & Impact: {input_outcome}
Structural Pattern: {input_pattern}

Per-candidate match/gap summary (5 dimensions each):
{candidate_summaries}
</input-event>
<reject-rules>
HARD REJECTION RULES — apply BEFORE picking closest. A candidate that violates any of these is DISQUALIFIED, even if its chain match looks tight. All examples below are TEMPLATED (angle brackets denote slot types).

REJECT-1: ALIAS — same event under different name.
Patterns:
– ⟨event design / planning / pilot phase⟩ disqualified for input that IS that design phase.
– ⟨ceremony's broadcast or telecast⟩ disqualified for input that IS the ceremony.
– ⟨war named by calendar A or duration⟩ disqualified for ⟨same war named by calendar B⟩.
– ⟨truncated form of input name⟩ disqualified for ⟨full input name⟩ when both refer to one specific incident.

REJECT-2: SUB-EVENT / SUB-PERIOD of the input.
Patterns:
– ⟨one named battle⟩ disqualified for input "⟨the multi-year war containing it⟩".
– ⟨key episode⟩ disqualified for input "⟨the decade-long movement containing it⟩".
– ⟨a stage / operation / sub-campaign⟩ disqualified for input "⟨the larger campaign that contains it⟩".
– ⟨named opening or final phase⟩ disqualified for input "⟨the encompassing event spanning that phase⟩".

REJECT-3: DIRECT SUPER-SET — input is a sub-event of the candidate.
Patterns:
– ⟨broader movement or campaign⟩ disqualified for input "⟨one named action within that movement⟩".
– ⟨encompassing war or era⟩ disqualified for input "⟨one operation / single dated incident within it⟩".

REJECT-4: GENERIC DESCRIPTOR of the same event type.
Patterns:
– ⟨category label⟩ ("Tournament", "Election", "Disaster", "Coup", "Industrial accident", "Ceremony") disqualified for input "⟨a specific dated/named instance⟩".
– ⟨unspecified-year placeholder⟩ ("Historical X", "Old X", "Past X") disqualified for input "⟨a specific named X with a date⟩".

REJECT-5: SCOPE-CLASS MISMATCH — candidate's temporal/magnitude class is at least one order of magnitude different from the input on a scale-dependent axis.
Patterns (DISQUALIFY when):

– input is a ⟨century-long era / multi-decade period⟩ AND candidate is a ⟨single-day / single-week named incident⟩.
– input is a ⟨multi-year movement / campaign⟩ AND candidate is a ⟨one-day protest / single dated action⟩.
– input has impact magnitude ≥1 OoM different from candidate on the relevant axis (casualty count, economic impact, geographic spread, role cardinality), AND the matched mechanism is scale-dependent.
EXCEPTION: if the candidate is a representative episode that captures the input era's DEFINING mechanism (e.g., a single decisive battle for a war), allow it under sub-event reasoning — but then it should also be checked under REJECT-2.

If a candidate matches ANY rejection rule, exclude it. If after rejection ZERO candidates remain, fall back to the highest-ranked non-rejected candidate even if mechanism match is weak.
</reject-rules>
<reasoning>
REASONING-FIRST PROCESS — work through these steps explicitly:

Step 1: For each of the {n_cands} candidates, walk the rejection rules above (REJECT-1 through REJECT-5). Mark each as ELIGIBLE or REJECTED (with reason).
Step 2: Among ELIGIBLE candidates, rank by:
(a) **MECHANISM SPECIFICITY:** PREFER candidates whose causal_chain match describes a SPECIFIC mechanism (concrete role + dynamic + outcome). DISPREFER candidates whose match is just a CATEGORY label (e.g., "both attacks", "both bubbles") even if "famous canonical".
(b) **SCALE MATCH** (within ∼1 order of magnitude, on these axes):
– Temporal class: input duration class ≈ candidate duration class (era / campaign / event must align, modulo REJECT-5 exception).
– Magnitude: casualty / economic / geographic-impact within ∼1 OoM.
– Role cardinality: 1-on-1 / 1-vs-many / many-vs-many should match.
A candidate with specific mechanism but mismatched scale by ≥1 OoM ranks BELOW one with both mechanism and scale match. Tight specific mechanism + matched scale > famous canonical pair.
Step 3: Pick top-ranked ELIGIBLE as closest.
Step 4: Decide `is_sufficient` (yes if mechanism match is reasonable AND scale roughly aligns; minor gaps OK).
Step 5: If not sufficient, identify common_gap and next_round_focus.
Step 6: Output JSON with reasoning field FIRST.

Now answer six questions for the JSON output:
Q1: Pick the CLOSEST ELIGIBLE candidate (after applying REJECT rules).
Q2: Is it SUFFICIENT? (mechanism reasonably matches AND scale roughly aligns).
Q3: REASONING — 2–3 sentences walking your rejection check + ranking + why the chosen candidate is best.
Q4: If insufficient, COMMON GAP across all candidates.
Q5: If insufficient, NEXT-ROUND ANGLE: (a) different mechanism family, (b) different scale/scope, (c) different outcome valence, (d) different role structure, (e) cross-domain shift.
Q6: List TOP-3 ELIGIBLE candidates (best first), excluding rejected ones.
</reasoning>
<output>
Output ONLY valid JSON, with reasoning FIRST so the LLM commits to the reasoning before the picks:
{
"reasoning": "Step 1 rejection check: ...
Step 2 ranking: ... Final pick justification.",
"closest_candidate": "⟨exact eligible name⟩",
"top_3": ["⟨best⟩", "⟨2nd⟩", "⟨3rd⟩"],
"is_sufficient": true OR false,
"common_gap": "..." (only if NOT sufficient; "" if sufficient),
"next_round_focus": "..." (only if NOT sufficient; "" if sufficient)
}
</output>

F. Rubric-based Evaluation for Historical Analogy Generation

We provide our new evaluation protocols for Li et al. (2025a).

Extraction prompt – 8-field structural decomposition

<task>
You are an event-analysis assistant. Given an event description, extract eight structural fields capturing its analogical structure (Sourati 2024 ARN narrative elements + causal-chain decomposition).
</task>
<fields>
1. ACTORS — principal agents by structural ROLE (e.g., "incumbent power", "rising challenger"), NOT proper names. 1–2 sentences.
2. RELATIONSHIPS — how actors relate (alliance, rivalry, dependence, hierarchy). 1–2 sentences.
3. ACTIONS — JSON array of 4–8 key actions in CAUSAL ORDER. Short phrases.
4. GOALS — what each actor aims to achieve. 1–2 sentences.
5. LOCATION — setting (geographical + institutional + domain). 1–2 sentences.
6. THEME — highest-order causal pattern in one phrase (e.g., "imperial overreach leads to quagmire").
7. CAUSAL_CHAIN — JSON array of ordered steps with VALENCE. Each step: `{"step": "⟨short phrase⟩", "valence": -1|0|1}` where $+1$ = building/positive trajectory, -1 = declining/negative, 0 = neutral. Use STRICTLY $\{-1, 0, 1\}$ — no ± 2.
8. CAUSAL_ELEMENTS — JSON array of 5–8 specific causal factors, snake_case short phrases. ONLY factors that play a CAUSAL role (drove or sustained the dynamic). NOT descriptions, dates, or proper names.
GOOD: `["imperial_overextension", "asymmetric_warfare", "domestic_antiwar_pressure"]`
BAD: `["nineteen sixty four", "Saigon", "war"]`
</fields>
<input>
Event:
{name}: {intro}
</input>
<output>
Output ONLY JSON, no other text:
```
{ "actors": "...", "relationships": "...", "actions": [...],
 "goals": "...", "location": "...", "theme": "...",
 "causal_chain": [{"step": "...", "valence": 1}, ...],
 "causal_elements": [...] }
```
</output>

Structural battery prompt – chain isomorphism + direction + mapping (Items 1–3)

<task>
You will rate STRUCTURAL alignment between two events on three sub-items, each 0–3. Focus on STRUCTURE — ignore proper names, dates, and domain.

EVENT A: actions {a_actions}; causal_chain {a_chain}; actors {a_actors}; relationships {a_rel}
EVENT B: actions {b_actions}; causal_chain {b_chain}; actors {b_actors}; relationships {b_rel}
</task>
<item-1: chain isomorphism>
Do A's action steps have functional counterparts in B, in roughly the same order?
0 = no correspondence (e.g., Tulip mania chain: speculation → bubble → crash vs Sermon on the Mount: gathering → teaching → dispersal).
1 = only ≤ 2 steps loosely correspond, rest unrelated (e.g., Russian Revolution vs French Revolution: both end in radical takeover but early steps don't map).
2 = most steps correspond with some misalignment or extras (e.g., Iraq War ↔ Vietnam War: invasion → insurgency → attrition → withdrawal align, but Iraq lacks Vietnam's "domino theory" framing step).
3 = clean step-by-step mapping (e.g., Iraq War ↔ Soviet-Afghan War: each step has a direct counterpart at similar position).
</item-1>
<item-2: direction consistency>
Does the trajectory shape match? Chains may have DIFFERENT NUMBERS OF STEPS or different granularity — compare overall trajectory SHAPE (where it rises/falls/turns), NOT exact step alignment.

Examples accommodating chain-length differences:
A: $[+1, +1, -1]$ (rise then fall, peak at 67%); B: $[+1, +1, +1, -1, -1]$ (rise then fall, peak at 60%) → **3** (isomorphic shape, granularity OK)
A: $[+1, +1, -1, -1]$ (rise-fall); B: $[-1, -1, +1, +1]$ (fall-rise) → **0** (opposite direction)
A: $[+1, +1, +1]$ (monotonic rise); B: $[-1, -1, -1]$ (monotonic fall) → **0**

A: $[+1, +1, +1, +1]$; B: $[+1, +1, +1] \rightarrow$ **3** (same monotonic, different length OK)
A: $[+1, +1, +1]$ (monotonic); B: $[+1, -1, +1, -1]$ (oscillation) $\rightarrow$ **1**
A: $[+1, +1, -1, -1]$ (rise-fall); B: $[+1, -1, -1, +1, -1]$ (turbulent) $\rightarrow$ **1**
A: $[+1, +1, -1, -1]$ (rise to $\sim$50%, fall); B: $[+1, +1, +1, -1]$ (rise to $\sim$75%, fall) $\rightarrow$ **2**

Score:
0 = opposite trajectory direction. 1 = different shape (monotonic vs oscillatory; rise-fall vs turbulent).
2 = roughly similar shape with some divergence in turning-point position. 3 = isomorphic trajectory.
</item-2>
<item-3: mapping consistency>
Can A's principal actors be mapped one-to-one to B's actors with parallel relationships? (Holyoak & Thagard 1989)
0 = roles cannot be mapped at all (e.g., Tulip mania merchants/speculators/buyers vs Battle of Hastings Normans/Saxons/king).
1 = ambiguous mapping (one A-role maps to multiple B-roles, or relationships cross).
2 = mapping works but at least one relationship diverges (e.g., A's "allies-turned-rivals" maps to B's "always rivals").
3 = clean one-to-one with parallel relationships throughout (e.g., Iraq War $\leftrightarrow$ Vietnam War: occupier $\leftrightarrow$ occupier, insurgents $\leftrightarrow$ insurgents, host government $\leftrightarrow$ host government).
</item-3>
<reasoning>
REASONING-FIRST PROCESS — work through these steps before assigning scores:
Step 1: Walk through Item 1 (chain). Map A's steps to B's. Decide level.
Step 2: Walk through Item 2 (direction). Compare valence trajectories. Decide level.
Step 3: Walk through Item 3 (mapping). Try the actor mapping table; check parallel relationships. Decide level.
Step 4: Output JSON. The "rationale" field MUST come FIRST (one sentence summarizing all three judgments). Numeric scores reflect the reasoning, not the other way around.
</reasoning>
<output>
```
{ "rationale": "one sentence summarizing the three judgments and the dominant
reason behind each score",
 "chain_isomorphism": 0, "direction_consistency": 0, "mapping_consistency": 0 }
```
</output>

Pragmatic battery prompt – overlap sufficiency + idiosyncrasy coverage (Items 4 and 6)

<task>
You will judge whether two events share IDIOSYNCRATIC causal elements vs only GENERIC ones.
</task>
<tagging-rule>
IDIOSYNCRATIC (I) = element names a SPECIFIC MECHANISM (specifies HOW or WHY). Short phrases like "imperial_overextension" count, because they specify a particular mechanism — not just a generic category.

GENERIC (G) = element is a CATEGORY LABEL (specifies WHAT but not HOW or WHY). "war", "election", "regime_change" apply to thousands of events without distinguishing mechanism.

Decision rule: Could this element distinguish one historical event from a generic example of its category? If YES $\rightarrow$ I. If it would apply to nearly any event of that type $\rightarrow$ G.

IDIOSYNCRATIC examples (specific mechanism — short phrasing OK):
`imperial_overextension`, `asymmetric_warfare`, `speculative_bubble`, `herd_valuation_beyond_fundamentals`, `ally_defection_under_attritional_pressure`, `logistical_overstretch_in_hostile_terrain`, `collective_boycott_as_diplomatic_weapon`, `nonviolent_mass_mobilization_against_single_party_rule`, `preemptive_strike_to_disrupt_alliance`, `domestic_antiwar_pressure_forcing_withdrawal`, `Old_Regime_dual_power_radicalization`.

GENERIC examples (bare category — almost any event has these):
`war`, `election`, `leadership_change`, `economic_decline`, `treaty_signing`, `international_pressure`, `military_defeat`, `domestic_protest`, `regime_collapse`, `religious_conflict`, `civil_unrest`, `diplomatic_crisis`.

Borderline:
`speculative_bubble` $\rightarrow$ I (specifies herd-valuation mechanism); `market_crash` $\rightarrow$ G (just outcome category).
`imperial_overextension` $\rightarrow$ I (specifies overextension as cause); `imperial_decline` $\rightarrow$ G (outcome only).

</tagging-rule>
<input>
EVENT A theme: {a_theme}; causal_elements: {a_elements}
EVENT B theme: {b_theme}; causal_elements: {b_elements}
</input>
<reasoning>
1. List shared elements that A and B genuinely share IN CONCEPT (identical wording NOT required — "imperial_overextension" and "overstretch_of_great_power" count as the same).
2. For each shared element: SPECIFIC MECHANISM (HOW/WHY) → I, or BARE CATEGORY (WHAT only) → G.
3. Score the two items below.
</reasoning>
<item-4: overlap sufficiency>
Are there enough IDIOSYNCRATIC shared elements?
0 = no shared elements at all.
1 = only GENERIC shared elements (just bare categories).
2 = exactly 1 idiosyncratic shared element.
3 = ≥2 idiosyncratic shared elements.
</item-4>
<item-6: idiosyncrasy coverage>
Of A's distinctive (specific-mechanism) causal elements, how many find a counterpart in B?
0 = none. 1 = ∼1/4. 2 = ∼1/2. 3 = most (≥3/4).
</item-6>
<output>
```
{ "shared_elements": [{"element": "...", "type": "I", "why": "..."}, ...],
 "overlap_sufficiency": 0, "idiosyncrasy_coverage": 0,
 "rationale": "one short sentence" }
```
</output>

Surface similarity prompt – attribute overlap (Item 5)

<task>
Score SURFACE similarity (concrete attribute overlap) between two events on a 0–3 scale. Surface = entity-level / attribute-level overlap (era, geography, named entities, domain). NOT structural pattern matching.

EVENT A: {a_name}; actors {a_actors}; location {a_loc}; goals {a_goals}
EVENT B: {b_name}; actors {b_actors}; location {b_loc}; goals {b_goals}
</task>
<categories>
Each of the 5 categories below counts AT MOST ONCE — do not double-count nested matches.

ERA — same time period within ∼30 years (count once even if same century AND decade AND year). E.g., "both early 20th C" = "both 1960s" = 1 attribute.

GEOGRAPHY — same continent OR country OR city (count once at most specific shared level). E.g., "both Europe" = "both France" = "both Paris" = 1 attribute.

NAMED ENTITIES — at least one specific named person/organization/country in BOTH. Count ONCE even if multiple overlap ("both involve US AND Britain" still = 1).

DOMAIN — same broad activity TYPE; pick MOST SPECIFIC SHARED level only, do NOT add nested layers. E.g., "both war" + "both naval war" + "both WWI naval war" = 1 attribute (the most specific shared = "WWI naval war").

OPERATION TYPE — specific operation (olympic boycott, speculative bubble, hostage standoff, ambush, scientific R&D program). ONLY count if different from / more specific than DOMAIN.
</categories>
<scoring>
Count how many of the 5 categories matched (each 0 or 1):
0 = 0 categories shared (e.g., 1980 Moscow Olympics ↔ Norman Conquest of England).
1 = exactly 1 category shared, typically only DOMAIN (e.g., 2008 GFC ↔ South Sea Bubble: only DOMAIN = financial bubble).
2 = 2 categories shared (e.g., Iraq War 2003 ↔ Vietnam War: DOMAIN = war + NAMED ENTITIES = US).
3 = ≥3 categories shared, OR sub-instance / near-self.

E.g., Iraq War 2003 ↔ Iraq War 2003 invasion phase (sub-instance — automatic 3).
E.g., Indo-Pakistani War 1965 ↔ Indo-Pakistani War 1971 (DOMAIN + GEOGRAPHY + NAMED ENTITIES + ERA = 4 → 3).
E.g., Yalta ↔ Potsdam Conference (DOMAIN + GEOGRAPHY + NAMED ENTITIES + ERA = 4 → 3).

CRITICAL: Be strict about anti-double-count. If two events are both "20th-century state-coercion famines in Eurasia", that's 3 categories (era + domain + geography), NOT 5.
</scoring>
<output>
`{ "shared_attributes": [...], "surface_similarity": 0, "rationale": "one short sentence" }`
</output>

System alignment prompt – canonical-form abstraction with reference levels (Item 8)

<task>
You will rate the structural alignment of two events' causal patterns on a 0–3 scale. You are given each event's actions, goals, and theme.

EVENT A: actions {a_actions}; goals {a_goals}; theme {a_theme}
EVENT B: actions {b_actions}; goals {b_goals}; theme {b_theme}
</task>
<step-1: canonical form>
Restate each event's CAUSAL PATTERN as a one-clause statement using ABSTRACT roles only (strip proper names, dates, domains). Use the schema:
"⟨role X⟩ ⟨action⟩ ⟨role Y⟩, resulting in ⟨outcome⟩, because of ⟨mechanism⟩".
</step-1>
<step-2: rate by reference pair>
LEVEL 0 — Different mechanism families. Almost no structural overlap.
Reference 0a:
A: "speculators inflate asset price, then panic sell when confidence breaks, collapsing market because herd behavior amplified valuation beyond fundamentals".
B: "religious leader challenges established doctrine, splits the community into rival factions, producing schism because ideological differences cannot be reconciled".
→ "speculation bubble" vs "religious schism" share NO mechanism. Different families.

LEVEL 1 — Same general PROCESS CATEGORY, but specific mechanism differs.
Reference 1a:
A: "expanding power launches sustained military offensive across multiple fronts, gets bogged down in attrition warfare, suffers strategic loss because coalition opponents organize sustained counter-resistance".
B: "ambitious power launches decisive surprise offensive, achieves rapid initial gains then collapses in attritional defense, suffers strategic loss because logistical overstretch overwhelms operational depth".
→ Both "military offensive that fails through attrition", but A's mechanism = COALITION COUNTER-RESISTANCE while B's = LOGISTICAL OVERSTRETCH.

Reference 1b:
A: "religious community participates in service-formation rite during commemorative period, reinforcing communal identity through shared ritual roles".
B: "religious community performs cyclical-renewal procession during commemorative period, reinforcing communal identity through symbolic re-enactment".
→ Both "religious commemorative ritual", but A's mechanism = SERVICE FORMATION while B's = SYMBOLIC RE-ENACTMENT.

LEVEL 2 — Same mechanism, but with noticeable differences in scope, role granularity, or outcome scale.
Reference 2a:
A: "expanding maritime empire launches piecemeal coercive intervention against fragmented island rulers, escalates from blockade to armed annexation, resulting in colonial control over a regional archipelago because local rulers fail to unite".
B: "expanding continental empire launches sustained coercive intervention against fragmented tribal confederations, escalates from punitive raids to full territorial conquest, resulting in colonial control over a continental landmass because local powers fail to coalesce".
→ Same mechanism ("piecemeal colonial conquest exploiting fragmentation"), different scope (Dutch in Lombok vs French in Algeria).

Reference 2b:
A: "speculators bid up prices of a novel agricultural commodity beyond use-value, market collapses in panic when confidence breaks, producing localized merchant losses because herd valuation exceeded what end-buyers would pay".
B: "speculators bid up prices of a government-backed corporate share beyond intrinsic value, market collapses in panic when confidence breaks, producing widespread elite financial losses and political fallout because herd valuation exceeded the company's actual revenue".
→ Same mechanism, different scope (tulips vs South Sea).

LEVEL 3 — Near-identical canonical forms: same role asymmetries, mechanism, outcome trajectory.
Reference 3a:
A: "weaker insurgent force conducts surprise attack on isolated detachment of stronger occupier, achieves limited tactical victory, converts to political momentum because symbolic win galvanizes broader coalition support against the occupier".
B: "weaker defender lures isolated detachment of stronger advancing army into ambush, achieves limited tactical victory, converts to political momentum because the symbolic win galvanizes broader resistance against the advancing power".
→ Near-identical structure (Battle of Princeton vs Battle of Cowpens).

Reference 3b:
A: "incumbent single-party authoritarian regime faces sustained nonviolent mass mobilization, security apparatus defects under public pressure, regime collapses without prolonged civil war because coercive capacity disintegrates".
B: "incumbent single-party authoritarian regime faces sustained nonviolent mass mobilization, security apparatus defects under public pressure, regime collapses with brief armed clash because coercive capacity disintegrates".
→ Near-identical with only minor difference (East Germany 1989 vs Romania 1989).
</step-2>
<output>
```
{ "a_canonical": "⟨role X⟩ ... resulting in ... because ...",
 "b_canonical": "⟨role X⟩ ... resulting in ... because ...",
 "system_alignment": 0,
 "rationale": "one short sentence explaining which reference level your A/B pair
most resembles" }
```
</output>

Novelty prompt – canonical vs novel pair detection (Item 9)

<task>
Judge whether the analogy pair (A, B) is NOVEL or CANONICAL on a 0–3 scale.

Definitions:
CANONICAL = pair appears frequently in history textbooks, classroom materials, news columns, or popular media (e.g., "Iraq War vs Vietnam"; "2008 GFC vs 1929 Depression"; "Russian Revolution vs French Revolution").
NOVEL = a non-obvious pair requiring structural insight to draw (cross-domain, cross-era, or cross-scale).

EVENT A: {a_name}; EVENT B: {b_name}
</task>
<reasoning>
Step 1: Briefly assess: would a college history textbook or popular-press history article likely treat (A, B) as a canonical comparison?
Step 2: Score on this scale:

0 = canonical pair, repeatedly drawn in standard textbooks.
E.g., 2008 GFC ↔ 1929 Great Depression; Iraq War 2003 ↔ Vietnam War; Russian Revolution ↔ French Revolution; WW2 ↔ WW1.

1 = well-known pair within history community, frequently drawn.
E.g., Iraq War 2003 ↔ Soviet-Afghan War; Pearl Harbor ↔ 9/11.

2 = less commonly drawn pair, requires some specialist knowledge.
E.g., Suez Crisis ↔ Falklands War; 1980 Olympics boycott ↔ 1976 Montreal African boycott.

3 = novel cross-domain or cross-scale pair requiring structural insight.
E.g., 2008 GFC ↔ Chernobyl Disaster; Apollo Program ↔ Manhattan Project (both: emergency cross-disciplinary R&D under deadline); French Revolution ↔ Iranian Revolution 1979.

</reasoning>
<output>
```
{ "is_canonical": true, "novelty": 0, "rationale": "one short sentence" }
```
</output>

Self-analogy 3-class prompt – IDENTITY / SIBLING_TRIVIAL / DISTINCT

<task>
Classify the relationship between historical events "{input_name}" and "{candidate_name}". Return ONE of three classes: **IDENTITY / SIBLING_TRIVIAL / DISTINCT**.
</task>
<class IDENTITY>
They refer to the SAME real-world event/thing, OR one is a strict part of the other. The candidate is NOT a valid analogy because it IS or is WITHIN the input.

(a) ALIAS — different names for the same specific event:
– ⟨event under codename⟩ ≡ ⟨same event under descriptive popular name⟩
– ⟨event under regional/local name⟩ ≡ ⟨same event under translated/global name⟩
– ⟨event under one academic framing⟩ ≡ ⟨same event under different academic/scholarly framing⟩ (e.g., "⟨Movement⟩ of YYYY" ≡ "⟨Same period⟩ Collapse of ⟨Container⟩")
– ⟨event under truncated form⟩ ≡ ⟨event under full form⟩
– ⟨event with date qualifier⟩ ≡ ⟨same event with no date⟩

(b) CATEGORY-SYNONYM — both names refer to the same category in standard usage (near-synonymous terms used interchangeably; same concept under formal vs. informal name).

(c) CONCEPT-LEVEL ALIAS WITH NO TOKEN OVERLAP — different names with no shared word but referring to the same historical moment or transformation (e.g., regional name ↔ international name; political label ↔ cultural label for the same period).

(d) CONTAINMENT — the candidate is a phase, sub-period, sub-event, sub-action, named operation, or specific incident WITHIN the input event's timeline/agenda/scope (or vice versa):
– Sub-event during input: candidate happened DURING input event's timeline.
– Member of movement: candidate is a specific named action within the input's broader movement/campaign.
– Phase/operation: candidate is a named operation within input's larger campaign.
</class IDENTITY>
<class SIBLING_TRIVIAL>
Distinct events, but they share a sibling-membership in one institution/series/actor's campaign such that their similarity is structurally trivial (explained by common cause, not by analogy):
– Numbered/recurring institutional series: both are members of an annual ceremony, sequential conference, world exposition, championship cycle, election cycle, etc.
– Same actor + same era + same action-type: same primary actor doing the same kind of action against similar targets in the same era (parallel campaigns of one program).
– Same series of meetings/treaties: both are installments of the same diplomatic/political series.

The events are independent, but their similarity comes from common cause/institution rather than from genuine structural analogy.
</class SIBLING_TRIVIAL>
<class DISTINCT>
Genuinely different events that deserve to be judged on their structural merits:
– Different events sharing only vocabulary (e.g., two wars whose names share a word but are distinct conflicts).
– Different events in the same domain at different times/places, NOT members of the same series.
– Cross-domain analogs (e.g., a war and a corporate takeover).
</class DISTINCT>
<reasoning>
Reason step by step BEFORE giving the final answer:
1. Identify the specific historical event each name refers to. Are they the same real-world referent under different names?
2. If different referents, check: is one a strict CONTAINMENT (phase / sub-event / sub-action / member-of-movement / sub-period) of the other?
3. If neither identity nor containment, check: are they SIBLING members of one institutional series / one actor's parallel campaign / one ceremony cycle?
4. Otherwise: DISTINCT.

Table 4. Seven failure modes of descriptive matching resolved by CANA decomposition.

Failure Mode	Descriptive Match (Li et al., 2025a)	CANA decomposition
Part-whole confusion	Arab Spring → Tunisian Rev.	→ 1989 E. Europe
Same-war echo	Expedition of 1000 → Volturno	→ Cuban Revolution
Self-analogy	Iraq War → 2003 Invasion	→ Soviet-Afghan War
Causal direction error	Battle of Bulge → Market Garden	→ Michael Offensive
Surface feature match	Caesar → Lincoln	→ Cromwell's death
Proximity bias	J. Brown's Raid → Bleeding Kansas	→ Easter Rising
Domain closure	Canudos → Siege of Balaler	→ Satsuma Rebellion

</reasoning>
<output>
Reasoning: ⟨2–3 sentences⟩
Answer: IDENTITY **or** SIBLING_TRIVIAL **or** DISTINCT
</output>

Seven failure modes resolved. Concretely, in Table 4, we present examples showing the seven systematic failure modes of descriptive matching via MDS decomposition, each resolved by CANA decomposition. It also echoes the importance of alignment via $M(E)$ as predicted by our theory.

F.1. Failure modes of descriptive matching