PolitNuggets: Benchmarking Agentic Discovery of Long-Tail Political Facts

Yifei Zhu
The University of Hong Kong
frankyifei@connect.hku.hk

Abstract

Large Reasoning Models (LRMs) embedded in agentic frameworks have transformed information retrieval from static, long-context question answering into open-ended exploration. Yet real-world use requires models to discover and synthesize "long-tail" facts from dispersed sources, a capability that remains under-evaluated. We introduce **PolitNuggets**, a multilingual benchmark for agentic information synthesis via constructing political biographies for 400 global elites, covering over 10,000 political facts. We standardize evaluation with an optimized Supervisor–Searcher multi-agent system and propose **FactNet**, an evidence-conditional protocol that scores discovery, fine-grained accuracy, and efficiency. Across models and settings, we find that current systems often struggle with fine-grained details, and vary substantially in efficiency. Finally, using benchmark diagnostics, we relate agent performance to underlying model capabilities, highlighting the importance of short-context extraction, multilingual robustness, and reliable tool use.

1 Introduction

Reasoning and synthesizing information within a given context is the defining capability of modern Large Reasoning Models (LRMs). The key framework can be called **Reasoning *in* Context**, where a model is *passively* provided a finite set of evidence and must extract or synthesize answers from it (Lewis et al., 2020; Guu et al., 2020). The rapid growth of context windows has enabled strong performance on long-document tasks (Shaham et al., 2023; An et al., 2024; Zhang et al., 2024; Bai et al., 2025; Vodrahalli et al., 2024; Yang et al., 2025b; Yen et al., 2025).

However, a new paradigm is emerging. By integrating LRMs into agentic frameworks equipped with retrieval tools, models can now actively explore, filter, and construct their own context from

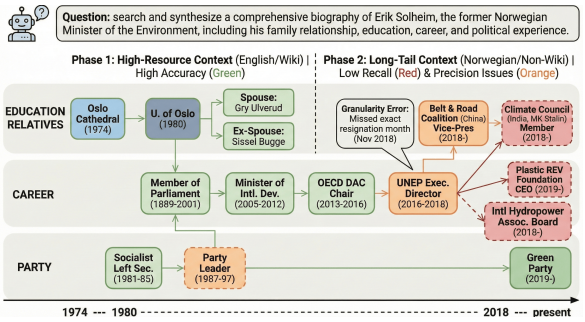

Figure 1: Agent performance heatmap on an example biography (Erik Solheim), illustrating the "head" vs. "long-tail" synthesis gap.

open-ended sources like the webpages and codebases (Nakano et al., 2021; Schick et al., 2023; Zhou et al., 2024). This unlocks a different layer of complexity: **Reasoning *through* Context**. Unlike the passive in-context setting, here the agent must navigate a potentially unbounded information space, making sequential decisions on *what* to read, *when* to stop, and *how* to synthesize fragmented evidence into a coherent whole (Wei et al., 2025).

While production systems like OpenAI Deep Research (OpenAI, 2025b) and Perplexity Deep Research (Perplexity AI, 2025) demonstrate the promise of this agentic paradigm, there remains a lack of rigorous benchmarks for Reasoning *through* Context under longitudinal synthesis demands. Many existing agentic evaluations emphasize short-horizon interactions or isolated fact retrieval (Yao et al., 2022; Mialon et al., 2024; Wei et al., 2025), and therefore under-measure the professional workflow of reconstructing a coherent narrative from scattered, disconnected, and sometimes contradictory sources. Further, few have linked the performance of a model's reasoning *through* context ability with the performance of a model's reasoning *in* context.

To bridge this gap, we introduce PolitNuggets, a benchmark grounded in a high-impact and re-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45012–45035
July 2-7, 2026

alistic task: the construction of political biographies. Wikipedia, while a triumph of collaborative human curation, exhibits systematic coverage gaps—particularly for non-US officials—and often lacks the fine-grained precision required for professional domains like, academic research or political consulting. PolitNuggets tests models' reasoning-through-context abilities by discovering the long-tail biography "nuggets" from the open web. This evaluation demands long-horizon reasoning, multi-language understanding, and reliable tool use. Our benchmark also characterizes a static corpus to evaluate models' reasoning-in-context ability.

Our evaluation of models within an agentic framework reveals that, although agents maintain high precision, they consistently struggle with recall in open-ended settings. We also observe a substantial performance degradation for Non-US entities (up to $\sim 40\%$ relative drop in F1 in some settings), highlighting a pronounced International Evidence Gap and demonstrating that multilingual robustness is a prerequisite for realistic use. We also connect the reasoning through context ability with the reasoning in context ability. Interestingly, the evaluation results reveal a **Long-Context Paradox**: strong long-context reading (Reasoning *in* Context) does not reliably predict end-to-end agent performance (Reasoning *through* Context); rather, success is driven by short-context reading precision, reliable tool use, and multi-language understanding.

1.1 Traversing a latent fact network

We conceptualize political biography reconstruction not as single-shot retrieval, but as traversing a latent fact network. Let a target biography induce a directed graph $G = (V, E)$, where nodes V are atomic "political nuggets" (e.g., *Minister of Defense, 2012–2015*) and edges E are latent temporal/causal links expressed in unstructured text (e.g., *"After resigning in 2015, she joined the World Bank"*). The agent starts from a seed (entity name and minimal metadata) and must recover the relevant subset of V by expanding along implicit edges discovered in documents.

This induces an optimization trilemma over correctness, coverage, and cost. Agents must maintain high precision (avoid unsupported events), high coverage (high recall over missing events in the long tail), and low efficiency cost (search steps/-tokens). This framing explains why naive RAG

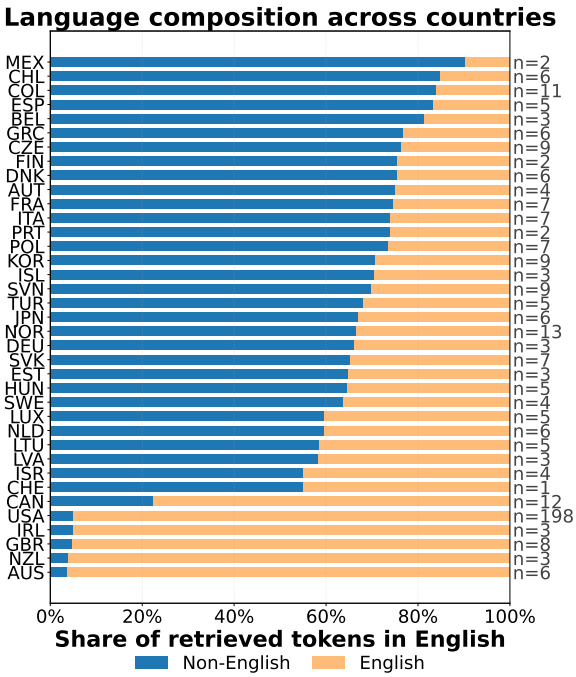

Figure 2: Language composition of retrieved evidence across countries. Bars show the share of retrieved tokens that are English vs. non-English; the right-side labels show the number of evaluated cases per country in our benchmark.

is insufficient: the missing long-tail nodes may be weakly connected, requiring multi-hop query reformulation and evidence accumulation.

PolitNuggets evaluates whether agents can approximate the full latent fact network while retaining the efficiency of strategic traversal. Strategic traversal jumps between salient nodes (low cost, but vulnerable to missing weakly connected phases if reasoning fails).

2 Benchmark & Task

The PolitNuggets benchmark evaluates agents on their ability to construct accurate, time-resolved career histories for 400 political elites sourced from global government directories.

2.1 Multilingual evidence in the wild

The evidence required to reconstruct political careers is inherently multilingual. Traversing a global biography is not merely a search problem: an agent must reason *through* multilingual context to decide what to read next, how to reformulate queries, and when a claim is sufficiently supported. To characterize the language composition of the documents an agent must consume, we compute (for each country) the fraction of retrieved evidence to-

kens that are in English versus non-English, based on the full set of pages and passages collected during the agentic experiments (Figure 2).

Our benchmark instances are drawn from the WhoGov dataset with a US and Non-US sampling design. We randomly sample 200 Non-US cabinet politicians from WhoGov (which provides names and basic metadata for over 58,000 global cabinet members from 1966 to 2023), and we also randomly sample 200 US legislators and senators. After preprocessing and filtering (e.g., ID matching), this yields the 400-entity evaluation set reflected in Figure 2.

2.2 Evaluation Levels: Event-Level vs. Attribute-Level

To disentangle an agent's ability to *find* relevant evidence from its ability to *extract fine-grained details*, we compute F1 at two levels of granularity, following standard slot-filling terminology.

1. **Event-Level F1 (Discovery):** Measures whether the agent correctly identifies the existence of a biographical event. A prediction is a true positive if the Role and Organization match the ground truth and the Year (Start/End) is correct. This primarily tests discovery (did the agent find the right nugget?).

2. **Attribute-Level F1 (Granularity):** Measures whether the agent can fill fine-grained attributes for an event (slot filling). A prediction matches only if the event-level criteria are met *and* the Start Month, End Month (within a 1-month tolerance), and Exact Official Title are correct. This primarily tests reading comprehension and schema compliance (did the agent read details correctly?).

The above slot structure (Role/Organization/-Date/Title) applies to career and party events; other event types use type-specific key fields (e.g., relation and name for relatives, institution and degree for education) with matching criteria adapted accordingly. Cross-lingual equivalence (e.g., Norwegian titles vs. English ground truth) is delegated to the evidence-conditional LLM judge rather than deterministic string normalization.

2.3 Experiment design and conditions

Model selection. To assess the current frontier of agentic information synthesis, we select models that jointly satisfy three constraints required by PolitNuggets: (i) Reasoning *in* Context (strong synthesis from a static context window), (ii) Reasoning *through* Context (robust tool use and multi-turn planning), and (iii) affordability/efficiency (enabling evaluation at the scale of hundreds of entities). As practical proxies, we prioritize models that score highly on OpenAI's Multi-Round Contextual Reasoning (MRCR) benchmark for long-context reasoning (OpenAI, 2025c) and on the Berkeley Function Calling Leaderboard (BFCL v3) for tool-use reliability (Patil et al., 2025), while favoring "Flash/Fast" variants or efficient open-weight models over prohibitively expensive frontier offerings. This yields our evaluated set: Grok-4-Fast (xAI, 2025), Gemini-2.5-Flash (Comanici et al., 2025), and Qwen-3 (80B/225B) (Yang et al., 2025a).

Task design. To disentangle *retrieval* capability from *discovery* capability, we evaluate models in two context conditions: with Wiki (Enhancement), where the agent is initialized with the target's existing Wikipedia text and must verify claims and fill missing gaps, and without Wiki (Reconstruction), where the agent starts from only the entity's name and must reconstruct the timeline from open-web sources (news archives, government gazettes) under a cold start.

3 Agentic System

3.1 Problem formalization

Let an entity e have a (latent) biography represented as a set of time-stamped events $G_e = \{v_1, \ldots, v_n\}$, where each $v_i = (r_i, o_i, t_i)$ denotes a Role r_i, Organization o_i, and a time interval t_i (e.g., start/end year or month). Let $W_e \subseteq G_e$ denote the subset covered by the entity's Wikipedia page (when present), and let P_e be the set of events predicted by an agent after interacting with the open web.

The agent executes a sequence of search queries $q_{1:T}$ under a policy $\pi(q_t \mid h_t)$, where h_t is the interaction history (retrieved snippets, intermediate notes, and partial timeline). Each query incurs a cost $c(q_t)$ (e.g., a search step and/or token usage), with a budget constraint $\sum_{t=1}^{T} c(q_t) \leq C$. The goal is to maximize coverage of missing biography events—i.e., high recall on $G_e \setminus W_e$—while remaining within budget:

$$\max_{\pi} \mathrm{E}\big[\mathrm{Recall}(P_e,\, G_e \backslash W_e)\big] \quad \text{s.t.} \quad \sum_{t=1}^{T} c(q_t) \leq C.$$

3.2 Architecture Details

We implement a standardized Supervisor–Searcher architecture with a clean tool interface to support long-horizon interaction while remaining operationally bounded (Figure 3).

1. **Supervisor:** Maintains global state via a running search summary and a to-do list. It decomposes the biography task into concrete search instructions for the Searcher and decides when to terminate the overall run (e.g., when marginal returns diminish or the step budget is reached).

2. **Searcher:** Executes search and browse/retrieve actions over unstructured web resources and returns targeted observations to the Supervisor. In addition to reporting observations, the Searcher can persist related chunks (source-linked evidence snippets) into an Archive. Keeping related records promotes detailed communication.

Finally, a specialized Coder agent maps the collected evidence into the strict JSON schema required for evaluation. In the final stage, we provide the Coder with both the Supervisor's report (summary + resolved to-do state) and the set of archived related chunks: the report provides global structure and resolved ambiguities, while the raw evidence supplies the fine-grained details needed for attribute filling.

An ablation study shows that adding Archive-backed evidence persistence yields a substantial gain (equivalently, removing the Archive drops Event-Level performance by $\Delta\mathrm{F1} \approx -0.05$), supporting memory as a core design choice (Appendix A.1.1).

Architecture vs. DeepResearch. Empirically, our agentic architecture produces a recall-oriented operating point: the best-performing setting in our system (Grok-4-Fast) achieves higher Event-Level recall than Gemini DeepResearch (powered by gemini 2.5 pro) in the With-wiki condition (US: 0.703 vs. 0.678; Non-US: 0.620 vs. 0.577), while Gemini DeepResearch is more precision-oriented (EventPrec US: 0.912 vs. 0.890; Non-US: 0.892 vs. 0.872; Appendix Table 4).

4 Evaluation Protocol

Standard exact-match metrics penalize agents for finding valid information that is absent from the ground truth (false positives). To address this, we employ the FactNet dynamic evaluation protocol. We report F1 at two levels of granularity: Event-Level F1 (discovery of the correct role/organization/year event) and Attribute-Level F1 (strict matching on fine-grained attributes such as start/end month and exact title, conditioned on a correct event).

4.1 Evaluation design

Let G_e be the evidence-verified biography nuggets for entity e, and let $W_e \subseteq G_e$ denote the subset covered by the entity's Wikipedia page. We construct G_e—the **Consolidated Ground Truth (CGT)**—incrementally from pooled evidence across agentic runs rather than from manual enumeration or Wikipedia. An initial batch of runs produces the seed set; as subsequent runs surface new candidate nuggets, each is verified against the proposing run's archived evidence using the Judge LRM and, if supported, added to G_e. All systems are scored against the final snapshot. Wikipedia is used only to define W_e (the coverage filter) and to support the With-Wiki condition. We validate CGT quality via manual timeline inspection (coverage) and human–LLM judge consistency audits plus independent fact-checking via Exa[1] (precision; Appendix A.1.1).

Our primary target is the novel set $G = G_e \setminus W_e$ (i.e., facts not already available on Wikipedia at evaluation time). Let P be the set of predicted nuggets produced by a system (agentic bio or LRM bio). We score predictions against a *dynamic novelty ground truth* G', initialized as G and expanded via novelty validation below to avoid penalizing supported discoveries missing from the curated set.

- **Novelty Validation (Dynamic Novelty CGT):** For any predicted nugget $p \in P$ such that $p \notin G$, we treat p as a candidate novel nugget and trigger verification. An external Judge LRM (gpt-5-mini) checks whether p is supported by the system's own evidence (source-linked passages in the Archive). If supported (and not Wikipedia-covered), p is added to G'; otherwise it remains a false positive. This yields a Dynamic Novelty CGT that credits verifiable new discoveries while maintaining evidence-grounded precision.

[1] https://exa.ai. Exa is an independent multilingual search backend used for audit.

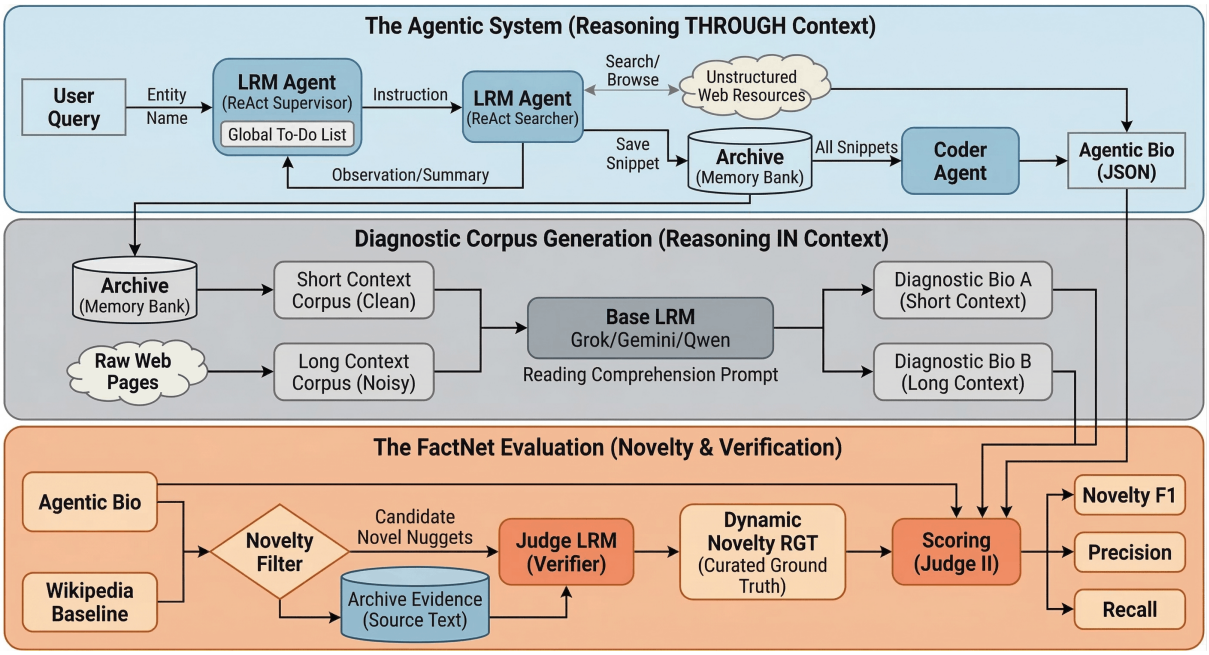

Figure 3: **The PolitNuggets Framework. (Top) Agentic system:** Supervisor+Searcher (+Archive) produces an Agentic Bio and the evidence corpora (Archive + retrieved pages). **(Middle) Long-context LRM baselines:** the Base LRM consumes these corpora to produce LRM bios (short-context from Archive; long-context from raw pages). **(Bottom) FactNet:** evaluates the bios with a dynamic novelty ground truth by filtering Wikipedia-covered facts and validating candidate novel nuggets against archived evidence.

- **Judge reliability checks:** We examined the consistency of this judge with human coders and an external search provider (Exa) via manual re-judging and independent fact checks (Appendix A.1.1).
- **F1 Score:** Calculated on the dynamic set G':

$$\text{Precision} = \frac{|P \cap G'|}{|P|},$$
$$\text{Recall} = \frac{|P \cap G'|}{|G'|},$$
$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

- **Efficiency Cost:** Measured as Average Search Steps per Entity and Total Token Usage. This quantifies the "cognitive effort" required to achieve a given F1 score.

4.2 Final evaluation

We evaluate two families of biographies produced from the same underlying evidence collection runs.

Agentic bios. Our agentic system produces Agentic Bios in two context conditions: With Wiki enhancement (4 models: Grok-4-Fast, Gemini-2.5-Flash, Qwen-3-225B, Qwen-3-80B) and Without Wiki reconstruction (2 models: Grok-4-Fast, Gemini-2.5-Flash), yielding 6 agentic bio types in total.

Long-context LRM bios (baselines). To quantify "Reasoning *in* Context" without agentic search, we ask each Base LRM to generate a biography directly from fixed evidence corpora produced by the Grok-4-Fast With-Wiki runs (the best-performing agentic setting), yielding 8 LRM bio types (4 models $\times$ 2 corpora): (i) a Short-context bio from the curated Archive (fine-grained, deduplicated evidence chunks; $\sim$30k tokens on average), and (ii) a Long-context bio from the concatenated Retrieved Web Pages (raw full documents from the same sessions; $\sim$300k tokens on average). This isolates improvements attributable to active planning, search, and evidence persistence (Reasoning *through* Context) versus what the Base LRM can achieve from a single static context window. Importantly, all LRMs are evaluated on the same fixed corpora.

5 Experimental Results

5.1 Main Performance Analysis

We analyze agent performance through the lens of a three-dimensional evaluation framework: Discovery (Event-Level F1), Granularity (Attribute-Level F1), and Efficiency (search steps/tokens).

Table 1: Main results. Performance is reported as F1 at two evaluation levels: Event-Level (discovery of role/organization/year) and Attribute-Level (slot filling of month-level dates and exact titles).

Context	Model	Region	EventF1	AttrF1
With wiki	Gemini DR	US	0.778	0.505
		Non-US	0.701	0.489
With wiki	Gemini	US	0.638	0.407
		Non-US	0.679	0.485
With wiki	grok-4 Fast	US	**0.768**	**0.501**
		Non-US	**0.712**	**0.475**
With wiki	qwen-225B	US	0.499	0.335
		Non-US	0.440	0.306
With wiki	qwen-80B	US	0.510	0.344
		Non-US	0.412	0.291
Without wiki	Gemini	US	0.671	0.439
		Non-US	0.618	0.468
Without wiki	grok-4 Fast	US	**0.766**	**0.506**
		Non-US	**0.708**	**0.475**

Note: The Without-Wiki condition is limited to Grok-4-Fast and Gemini-2.5-Flash because cold-start reconstruction retrieves substantially more documents, which can exceed Qwen's maximum context window (256k tokens), leading to unstable runs.

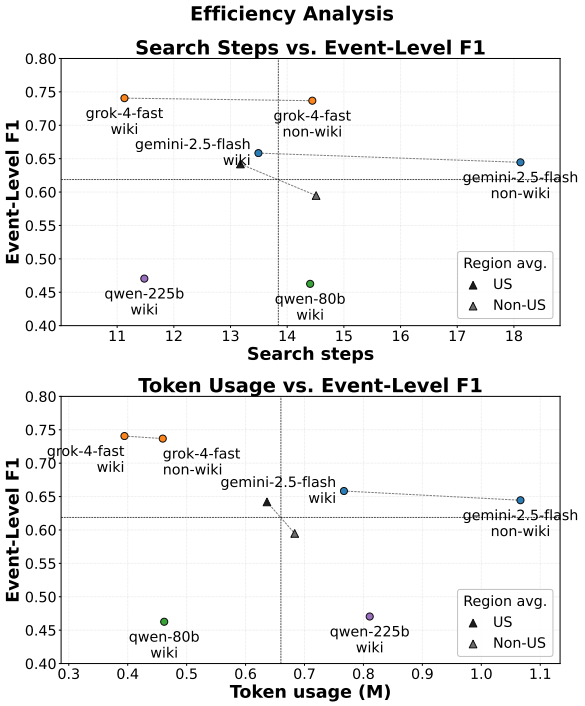

Figure 4: Efficiency Analysis: Search steps vs. F1 (top) and Token usage vs. F1 (bottom). Grok-4-Fast occupies the efficient frontier (top-left), achieving high F1 with minimal steps/tokens—a form of superior cognitive economy. Without Wikipedia context, Gemini maintains similar accuracy but requires substantially higher search volume (long dashed lines), relying on more search rather than improved reasoning efficiency.

Table 1 presents the comprehensive performance across all experimental conditions. Grok-4-Fast is the strongest model across both evaluation levels and both contexts, while also using fewer search steps. With Wiki, Grok-4-Fast achieves the best Event-Level F1 (US: 0.768; Non-US: 0.712) and the best Attribute-Level F1 (US: 0.501; Non-US: 0.475) at 11.1 steps on average; without Wiki it remains strong (Event-Level US/Non-US: 0.766/0.708; Attribute-Level US/Non-US: 0.506/0.475) with 14.5 steps. In contrast, Gemini exhibits comparable F1 in some settings but at substantially higher cost in cold-start reconstruction (18.1 steps), and Qwen variants trail in both Event-Level discovery and Attribute-Level slot filling.

Finding 1: Discovery and granularity remain unsolved—and the gap is driven by recall, not precision. Performance drops sharply when moving from Event-Level to Attribute-Level evaluation, and even Event-Level discovery is far from saturated. For example, Grok-4-Fast drops from 0.768 to 0.501 F1 (US), showing that extracting month-level dates and exact titles remains difficult. Decomposing performance reveals that this shortfall is primarily a recall/coverage problem rather than a precision problem: precision remains consistently high while recall is substantially lower and deteriorates further at the Attribute-Level (Appendix Table 4). In other words, agents are largely conservative—they tend to miss weakly connected long-tail events and attributes rather than fabricate unsupported ones. This precision–recall shape mirrors the passive vs. active split: models are often strong at Reasoning *in* Context once a relevant snippet is found, but fail at Reasoning *through* Context when they must autonomously discover that snippet in the first place (Section 6 and Table 3).

Finding 2: The International Evidence Gap. We observe a consistent performance degradation for Non-US entities across most models. While Gemini maintains parity in Event-Level F1 with Wiki (0.638 US vs. 0.679 Non-US), Qwen-225B drops from 0.499 (US) to 0.440 (Non-US) at the Event-Level, and performance also degrades at the Attribute-Level across settings. This highlights a structural bias: lower availability of English-language sources and higher complexity in parsing non-US government archives significantly hamper agentic synthesis.

5.2 Efficiency Analysis: The Pareto Frontier

To visualize the trade-off between performance and computational cost, we plot the Efficiency Pareto Frontier in Figure 4. The dashed vertical and horizontal reference lines denote the average cost (steps/tokens) and the average F1, splitting the plane into four quadrants. The desirable "nice zone" is the top-left quadrant (above-average F1 with below-average cost), while the bottom-right quadrant corresponds to below-average F1 with above-average cost.

Two robustness patterns emerge. First, removing Wikipedia context substantially increases cost (points shift rightward to higher steps/tokens), yet accuracy changes modestly. We interpret this "Wiki removal" setting as a stress test of agentic robustness under prolonged trajectories: even as reasoning and search steps rise, F1 does not collapse, suggesting the framework can sustain longer-horizon interaction without severe degradation once it accumulates sufficient evidence. Second, Non-US settings tend to be less efficient than US settings, with many Non-US points shifting toward higher cost and/or lower F1, consistent with a multilingual evidence burden and more fragmented source ecosystems.

Finding 3: Wiki removal reveals efficiency gap. Figure 4 shows that removing Wiki context consistently shifts points rightward (higher steps/tokens), moving runs out of the top-left "nice zone.". Across models, we observe a clear cognitive-economy gap: Grok typically achieves comparable F1 with fewer steps, while Gemini more often substitutes search volume for reasoning precision (a "brute force" strategy). Taken together, these results suggest that the core remaining challenge is not simply "more thinking," but more efficient retrieval and search strategy: better query planning and evidence targeting are required to achieve high coverage without paying a large cost increase. **Statistical significance.** The key efficiency gaps highlighted here (e.g., Wiki removal increasing steps/tokens for Gemini and Grok) have bootstrap 95% confidence intervals for mean deltas that exclude 0; see Appendix A.2.

6 LRM Analysis: Bridging Passive Reasoning and Active Discovery

Having established the agentic performance benchmarks in Section 1 and the controlled "Reasoning *in* Context" baselines in Section 4.2, we now investigate the fundamental drivers of success in longitudinal information synthesis. Using the short/long-context corpora defined in Section 4.2 and the FactNet protocol in Section 4.1, we decompose performance into six diagnostic dimensions: Short-Context Extraction, Long-Context Recall, the Long–Short Gap, Parametric Knowledge, Multilingual Robustness, and Tool-Use Reliability (BFCL). This analysis aims to bridge the gap between traditional "Reasoning *in* Context" benchmarks and the emerging "Reasoning *through* Context" paradigm. For completeness, we report the full LRM baseline results table in Appendix A.3 (Table 3).

6.1 The primacy of short-context extraction

We observe that a model's ability to extract facts from a clean, short context—the curated Archive—is strongly predictive of end-to-end agentic performance (Figure 5a). This suggests a "last-mile" bottleneck: if the model cannot reliably parse and structure high-quality evidence it has already found, additional searching cannot recover the loss, and end-to-end synthesis degrades primarily through missed events and attributes.

6.2 The decoupling of long-context recall

Crucially, we find that passive long-context recall on the noisy, full-document baseline is a weak predictor of agentic success. In particular, models that dominate end-to-end discovery do not necessarily outperform peers on static long-context extraction. This validates an *agentic hypothesis*: episodic search that iteratively curates small, high-quality contexts can outperform reliance on massive context windows alone, effectively bypassing the limitations of "Reasoning *in* Context" under noise.

An unintuitive finding is that the Long–Short gap is not a reliable proxy for agentic success. Degradation from curated short contexts (Archive) to noisy long contexts (full retrieved pages) does not consistently track end-to-end agent F1 (Figure 5b–c). One plausible explanation is a *training dilemma*. If so, "better long-context" and "smaller long–short degradation" may not co-exist monotonically under realistic budget and model-structure constraints.

6.3 The multilingual reasoning barrier

The "International Evidence Gap" observed in our main results is structurally explained by a multilingual reasoning barrier (Figure 5e). Models that

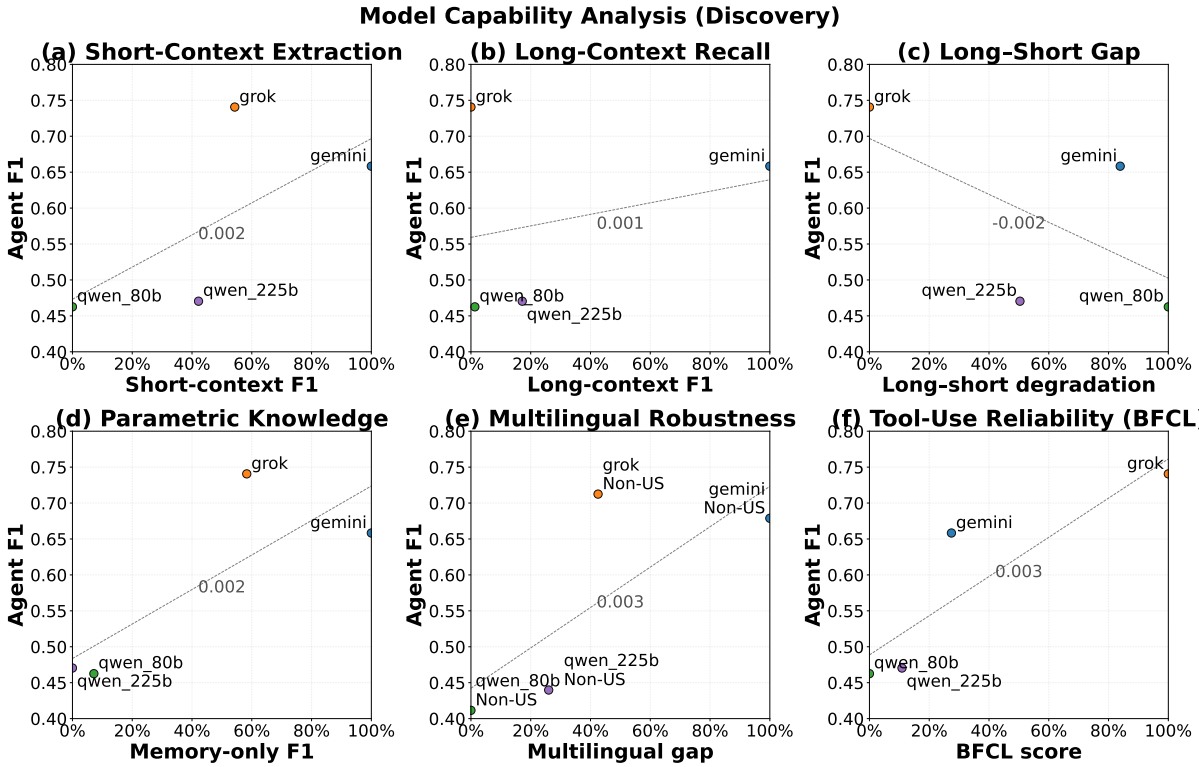

Figure 5: Model capability analysis (Event-Level). Each panel plots a normalized capability score (x-axis) against end-to-end Agent F1 (y-axis). **(a)** Short-Context Extraction, **(b)** Long-Context Recall, **(c)** Long–Short Gap, **(d)** Parametric Knowledge (closed-book), **(e)** Multilingual Robustness, **(f)** Tool-Use Reliability (BFCL). A positive trend indicates that the capability predicts agentic success.

exhibit larger degradation when extracting from non-English evidence chunks also underperform on Non-US entities. This indicates that global longitudinal synthesis is bottlenecked not only by retrieving foreign-language documents, but by reasoning over them with comparable fidelity to English evidence.

6.4 Parametric knowledge and tool reliability as scaffold

Finally, we observe a positive relationship between a model's closed-book (no-evidence) biography ability and its capacity to discover missing facts (Figure 5d,f). Parametric knowledge appears to act as a semantic scaffold: it supports entity disambiguation, improves query formulation, and helps the agent recognize valuable nuggets when encountered in the wild. Importantly, this semantic scaffold only helps end-to-end if the model can *reliably act on it*: tool-use reliability (BFCL) complements parametric knowledge by reducing brittle failures in search/browse execution, enabling consistent multi-step query reformulation, and translating high-level intent into stable tool calls.

7 Related Works

Reasoning *in* context: from needle-in-a-haystack to graph reasoning. A rich line of long-context benchmarks has progressively raised the difficulty of passive evidence extraction. Early "needle-in-a-haystack" setups (e.g., HELMET (Yen et al., 2025)) probe whether models can locate a single target fact in a long context. Subsequent benchmarks extend this to multiple needles and multi-round contextual reasoning (MRCR (OpenAI, 2025c)), and further to structured reasoning over latent or explicit graphs (Michelangelo (Vodrahalli et al., 2024), GraphWalks (OpenAI, 2025a)). Closest to our setting, LongBioBench (Yang et al., 2025b) uses controlled synthetic biographies to examine long-context understanding, reasoning, and trustworthy generation. PolitNuggets progresses from this lineage by shifting the locus of difficulty from reasoning *in* a given context to reasoning *through* context—where the agent must actively discover, filter, and synthesize evidence from the open web—and by grounding evaluation in real-world, multilingual political biographies rather than synthetic text.

Evaluating reasoning *through* context: Evaluation of tool-augmented reasoning spans static multi-hop QA datasets such as MuSiQue (Trivedi et al., 2022) and general agentic benchmarks such as GAIA (Mialon et al., 2024) that assess fundamental tool proficiency. More recently, benchmarks including WebSailor (Li et al., 2025) and BrowseComp (Wei et al., 2025) emphasize verifying retrieved information in open environments, often following a "hard-to-find, easy-to-verify" paradigm (e.g., identifying a specific paper from indirect cues) (Wei et al., 2025). DeepResearch Bench (Du et al., 2025) pushes toward realistic research workflows with expert-crafted tasks, but this style of evaluation can be expensive and remains sensitive to verification quality. PolitNuggets builds on this line while moving beyond isolated fact lookup to multi-faceted biography synthesis, and provides a scalable evaluation protocol for multilingual discovery, addressing a critical gap in global agentic information retrieval.

8 Conclusion

PolitNuggets provides a rigorous assessment of agentic information synthesis in the wild, targeting the gap between Reasoning *in* Context (fixed evidence) and Reasoning *through* Context (tool-driven discovery). We introduce a scalable benchmark of political biography construction for 400 global elites and evaluate systems with FactNet, an evidence-conditional protocol that measures discovery, fine-grained accuracy, and efficiency while validating candidate nuggets against retrieved evidence. Across models and settings, we find that precision is generally high but performance is recall- and efficiency-limited, with Wikipedia removal substantially increasing search cost, and we observe a pronounced International Evidence Gap on Non-US entities. Ablations show that evidence persistence improves end-to-end outcomes, and our diagnostics highlight a "Long-Context Paradox": strong long-context reading does not reliably predict agentic success, which is instead driven by short-context extraction, multilingual robustness, and reliable tool use. We hope PolitNuggets and its released artifacts support reproducible progress on evaluating and improving real-world agentic systems.

Acknowledgments

We thank Songpo Yang, Xiao Liu, Jiangnan Zhu, and Junyan Jiang for insightful discussions around this project. We also thank the anonymous reviewers for their constructive feedback.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longbench v2: towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and 1 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: a comprehensive benchmark for deep research agents. *Preprint*, arXiv:2506.11763.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. Websailor: navigating super-human reasoning for web agent. *Preprint*, arXiv:2507.02592.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom.

2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations (ICLR)*. Accessed: 2026-01-06.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2025a. GraphWalks: a multi-hop long-context graph reasoning benchmark. https://huggingface.co/datasets/openai/graphwalks. Released with GPT-4.1; results reported in the GPT-4.1 blog post.

OpenAI. 2025b. Introducing deep research. https://openai.com/index/introducing-deep-research/. OpenAI release, published February 2, 2025.

OpenAI. 2025c. OpenAI-MRCR (Multi-Round Coreference). https://huggingface.co/datasets/openai/mrcr. OpenAI dataset/eval introduced in the GPT-4.1 release on April 14, 2025; results used in this paper are provided in `final_eval/mrcr_score.csv`.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The berkeley function calling leaderboard (bfcl): from tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*. OpenReview: accessed 2026-01-05.

Perplexity AI. 2025. Introducing Perplexity Deep Research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research. Perplexity AI Blog.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, and 5 others. 2024. Michelangelo: long context evaluations beyond haystacks via latent structure queries. *arXiv*, (arXiv:2409.12640). Version 2.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: a simple yet challenging benchmark for browsing agents. *Preprint*, arXiv:2504.12516.

xAI. 2025. Grok 4 Fast Model Card. https://x.ai/model-card/grok-4-fast. XAI Technical Report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yijun Yang, Zeyu Huang, Wenhao Zhu, Zihan Qiu, Fei Yuan, Jeff Z. Pan, and Ivan Titov. 2025b. A controllable examination for long-context language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. Helmet: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*. OpenReview: https://openreview.net/forum?id=293V3bJbmE.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞fty Bench: Extending Long Context Evaluation Beyond 100K Tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

9 Limitations

First, due to budget constraints and practical model selection, we do not evaluate the largest and most expensive frontier-scale models. Such models may

reveal a clearer connection (or a different relationship) between Reasoning *in* Context and Reasoning *through* Context.

Second, although we provide cached results for reproducibility, benchmark outcomes may still shift over time due to changes in the underlying search engine and the evolving web (ranking drift, content updates, and availability).

Third, our static-context LRM baselines are constructed from evidence corpora produced by agentic collection runs. This yields a controlled comparison, but it may not fully represent long-context performance under independently collected evidence. Thus, we can not conclude that reasoning through context is better than reasoning in context in this research.

10 Ethical Considerations

PolitNuggets is constructed from human-related information available in the public domain (e.g., Wikipedia, official government pages, and public news/biographical sources). We adhere to fair-use principles for research and release only cached materials necessary for replication. We do not intentionally collect or disclose private or sensitive personal information beyond what is already publicly available, and we do not include any leaked/private datasets.

11 Potential Risks

The agentic biography-construction techniques evaluated here could in principle be repurposed to profile private individuals or non-public figures; we therefore restrict our benchmark to public officials whose career information is a matter of public record. Model-generated biographies can also contain factual errors that, if redistributed uncritically, could harm the subjects or propagate misinformation; we mitigate this by scoring only against evidence-verified nuggets and by releasing the cached source pages so that each claim can be audited. Finally, the International Evidence Gap we document indicates that naive deployment of such systems may disproportionately under-represent non-US and non-English political figures, and this limitation should be explicitly disclosed in any downstream use.

12 AI Usage

We used AI tools to assist with (i) implementation and refactoring of the evaluation and analysis code, (ii) plotting and figure formatting, and (iii) proofreading and minor style edits of the manuscript. All experimental design decisions, evaluations, and reported results were reviewed by the authors, and we validated code changes by re-running the pipeline to reproduce tables and figures.

A Additional Analysis

A.1 Experiment Details

A.1.1 Deliverables

To support replication, we release a code repository[2] containing the Supervisor–Searcher agentic pipeline, together with a data release[3] containing the following three artifacts:

1. **Consolidated Ground Truth (CGT).** The final pooled, evidence-verified biography nuggets for all 400 entities (including the Wikipedia-coverage filter W_e), which define the evaluation target G and the dynamic novelty set G'.

2. **Cached webpages.** The raw retrieved web pages collected during our agentic runs, fixing the search snapshot used for all reported numbers and enabling offline re-evaluation.

3. **LRM evaluation package.** A curated static-context dataset (Archive-style short context and long-context corpora derived from the cached pages) for evaluating long-context biography extraction without interactive search, enabling controlled comparison of "Reasoning *in* Context" across models.

All LRM-baseline and FactNet evaluation procedures are fully specified by the prompts in Appendix A.5 together with the released artifacts above, allowing reassembly without a dedicated code release.

Infrastructure, tools, and cost accounting. We implemented the full Supervisor–Searcher pipeline in `langgraph`. For LLM inference, we used OpenRouter as the API provider and recorded token usage using OpenRouter's standardized token accounting. For web search, we used the Serper API; for page retrieval, we used the scrawling service by Jina and Exa. To maximize robustness at scale, we used multiple layers of retry/backoff to

[2] https://github.com/yifeifrank/poli_searcher

[3] https://huggingface.co/datasets/frankyifei/politnuggets

ensure successful search and retrieval. Across the full experimental campaign (including development/testing), we issued approximately 300k searches (about $300 in search cost) and retrieved pages at an additional cost of about $50 (including free-tier, in-limit usage). Overall, the total third-party API spend for the project (including LLM APIs, search, and retrieval) was approximately $3,750.

Budget controls and termination criteria. We enforced two complementary termination criteria to bound the system's budget. First, the Supervisor maintains a to-do list and allocates a bounded amount of *dedicated research* per item: for each to-do item, the Searcher is allowed at most three focused search–retrieve attempts; if the item remains unresolved after three attempts, the Supervisor abandons that branch and proceeds to other items to avoid pathological loops. Second, we impose a hard-coded global cap of 100 LLM calls per run to bound worst-case cost and latency.

Experimental timeline and operational exclusion. Our primary data collection was conducted in September 2025 (Gemini and Qwen families), and we added Grok-4-Fast in November 2025 given its strong tool-use reliability and direct relevance to real-world agentic deployments. We also attempted to evaluate GPT-5-Mini, but in preliminary runs it frequently failed to terminate within the maximum conversation times budget ($T = 100$), exhibiting repetitive search loops and unfinished trajectories; we therefore exclude it from the final comparison to preserve integrity of the evaluation. However, due to it's strong long context performance and affordability, we used as the judge LRMs in this article.

Consistency and validity of the LLM judge. To assess the reliability of our evidence-conditional LLM judge, we manually re-judged the final scores for 40 randomly selected officials and compared them against the LLM judge outputs. The human and LLM judgments are consistent, with a correlation of 0.87. As an additional validity check, we used Exa to further fact-check a random sample of 100 officials: out of 3,240 validated nuggets, we identified 23 false positives, corresponding to an inaccuracy rate of $\approx 0.71\%$. The manual re-judging was performed by four student annotators recruited through the authors' university network. Each annotator was compensated at HKD 70 per hour, which is the prevailing rate for comparable

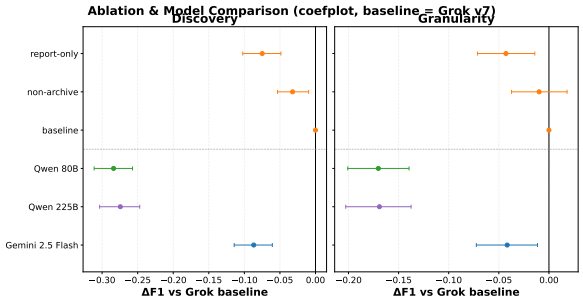

Figure 6: Architectural Ablation (Coefplot). Removing the Archive ("non-archive") significantly degrades performance (ΔF1 ≈ -0.05), confirming that raw evidence persistence is crucial for longitudinal synthesis.

student research assistance in Hong Kong. Annotators were informed in writing about the purpose of the task (validating LLM-generated biography assessments of public political figures), the use of their labels (aggregate statistics only, no personal data collected). The instruction given are similar to prompts used in research.

Architectural ablation: the necessity of memory. To validate the Archive memory mechanism added on top of the base Supervisor–Searcher loop, we conduct ablation studies on the Grok-4 baseline (Figure 6). We compare the full system against a No-Archive variant (where evidence persistence is disabled and the Supervisor relies only on the Searcher's summaries) and a Report-Only variant.

The results are unequivocal. The coefficient for non-archive is significantly negative ($\beta \approx -0.05$ at the Event-Level), validating our hypothesis: summaries are lossy. Without the Archive Tool to persist raw evidence chunks, the agent suffers from "Contextual Amnesia," hallucinating connections or failing to link dependent facts.

A.2 Statistical significance

To support the claims about setting and region gaps, we report distributional summaries and bootstrap confidence intervals for the underlying per-entity metrics. Table 2 provides key contrasts (mean deltas with bootstrap 95% CIs; "CI excludes 0" indicates statistical significance at the 5% level).

In table 2, each row corresponds to one *difference* computed over entities. **comparison** specifies the direction: (`Non-US minus US`) is $\mathrm{E}[\text{Non-US}] - \mathrm{E}[\text{US}]$ within a fixed context/model; (`Without wiki minus With wiki`) is $\mathrm{E}[\text{Without}] - \mathrm{E}[\text{With}]$ for the same model. **level** indicates whether the outcome is **Discovery** (event-level), **Granular-**

ity (attribute-level), or **Efficiency** (cost). **metric** is the quantity being compared (`f1`, `steps`, or `tokens_total`). **n_us** and **n_non_us** are the effective sample sizes (completed entities) for the two groups, and **mean_us / mean_non_us** are the corresponding sample means. **delta** is the mean difference in the direction defined by **comparison**. **ci95_low** and **ci95_high** are the bootstrap 95% confidence interval bounds for **delta**, and **ci95_excludes_0** is `True` when the CI does not cross 0 (significant at $\alpha = 0.05$). **cohen_d** reports a standardized effect size (sign follows **delta**). We use unpaired bootstrap resampling for `Non-US minus US` (different entity sets) and paired bootstrap for `Without wiki minus With wiki` when comparing the same entities across settings.

A.3 Full experiment results

A.3.1 LRM baseline results

LRM baseline findings. First, both Short and Long LRM bios underperform the best agentic setting (e.g., Grok-4-Fast With Wiki: EventF1 0.768 / AttrF1 0.501), despite operating over evidence collected from the same sessions. Second, Long-context performance is consistently worse than Short-context performance; for Grok-4-Fast (US, EventF1) the drop is $(0.626 - 0.538) \times 100/0.626 \approx 14.1\%$, reflecting degradation under long, noisy contexts. Third, the Memory-only baseline is uniformly low, suggesting that while memory leakage exists, it is not the deterministic driver of success in this task compared to evidence-grounded extraction.

A.3.2 Precision and recall breakdown

Breakdown. Table 4 confirms that the dominant gap is coverage rather than factuality: across models, EventPrec is consistently high while EventRec is substantially lower, and the drop is even sharper at the Attribute-Level (month/title matching). Notably, Gemini DeepResearch exhibits the highest precision in the With-wiki setting (EventPrec US/Non-US: 0.912/0.892) but lower recall than our best agentic model (EventRec US/Non-US: 0.678/0.577 vs. Grok-4-Fast 0.703/0.620), indicating a more conservative operating point. This precision–recall shape supports our framing that longitudinal synthesis failures are primarily due to missed weakly connected long-tail events, especially for Non-US entities.

A.3.3 Model capability analysis (Attribute-Level)

Figure 7 presents the diagnostic analysis for the Attribute-Level evaluation. The trends largely mirror the Event-Level setting, though the correlations are noisier due to the overall lower performance ceiling under strict month-level matching.

Table 2: Statistical evidence for key differences. We report mean deltas with bootstrap 95% confidence intervals (CIs) computed over entities; "CI excludes 0" indicates significance.

comparison	level	context	model	metric	n_us	n_non_us	mean_us	mean_non_us	delta	ci95_low	ci95_high	ci95_excludes_0	cohen_d
Non-US minus US	Discovery	With wiki	grok-4 Fast	f1	193	189	0.7682	0.7125	-0.0557	-0.0958	-0.0163	True	-0.2852
Non-US minus US	Granularity	With wiki	grok-4 Fast	f1	193	189	0.5011	0.4755	-0.0256	-0.0725	0.0213	False	-0.1105
Non-US minus US	Discovery	Without wiki	grok-4 Fast	f1	185	186	0.7656	0.7081	-0.0575	-0.0948	-0.0188	True	-0.3040
Non-US minus US	Granularity	Without wiki	grok-4 Fast	f1	185	186	0.5056	0.4748	-0.0307	-0.0724	0.0137	False	-0.1411
Non-US minus US	Discovery	With wiki	Gemini	f1	196	194	0.6380	0.6788	0.0408	-0.0023	0.0834	False	0.1843
Non-US minus US	Granularity	With wiki	Gemini	f1	196	194	0.4067	0.4854	0.0787	0.0322	0.1266	True	0.3399
Non-US minus US	Discovery	Without wiki	Gemini	f1	195	191	0.6706	0.6179	-0.0526	-0.0938	-0.0111	True	-0.2521
Non-US minus US	Granularity	Without wiki	Gemini	f1	195	191	0.4393	0.4678	0.0285	-0.0163	0.0726	False	0.1270
Non-US minus US	Discovery	With wiki	qwen-225B	f1	191	179	0.4991	0.4398	-0.0593	-0.1087	-0.0107	True	-0.2463
Non-US minus US	Granularity	With wiki	qwen-225B	f1	191	179	0.3350	0.3056	-0.0295	-0.0733	0.0134	False	-0.1352
Non-US minus US	Discovery	With wiki	qwen-80B	f1	191	179	0.5104	0.4115	-0.0989	-0.1453	-0.0517	True	-0.4362
Non-US minus US	Granularity	With wiki	qwen-80B	f1	191	179	0.3439	0.2910	-0.0529	-0.0985	-0.0080	True	-0.2365
Without wiki minus With wiki	Efficiency	Gemini	Gemini	steps	394	394	13.5330	18.0426	4.5096	3.0324	5.9313	True	0.4029
Without wiki minus With wiki	Efficiency	Gemini	Gemini	tokens_total	394	394	770151.3452	1062534.4629	292383.1178	143694.3691	449362.6692	True	0.2610
Without wiki minus With wiki	Efficiency	grok-4 Fast	grok-4 Fast	steps	372	372	11.1694	14.5188	3.3495	2.3144	4.3441	True	0.4469
Without wiki minus With wiki	Efficiency	grok-4 Fast	grok-4 Fast	tokens_total	372	372	394522.1747	461226.9086	66704.7339	32969.7364	99277.5562	True	0.2287

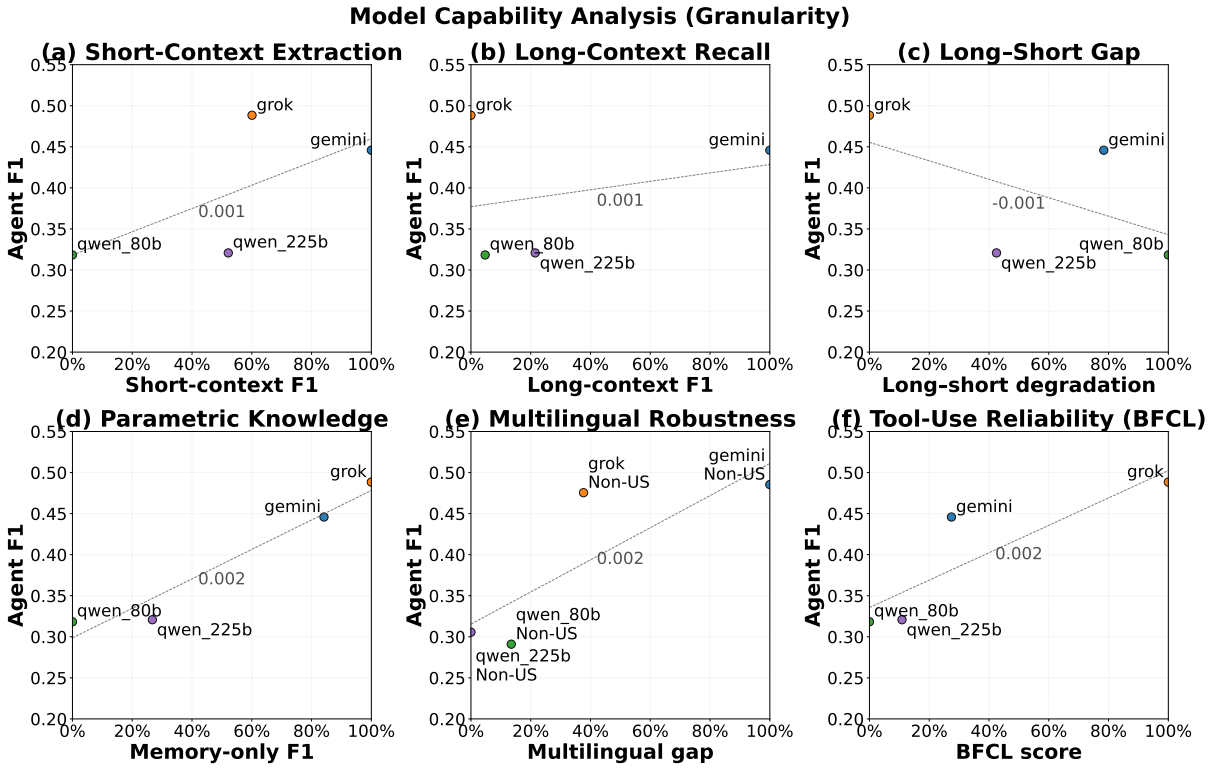

Figure 7: Model Capability Analysis (Attribute-Level). The same 2×3 diagnostic grid as Figure 5, with panels (a)–(f), for the Attribute-Level evaluation.

Table 3: LRM baseline results (Reasoning *in* Context). We report F1 for biographies generated directly from three static contexts: **Short** (Archive), **Long** (raw retrieved pages), and **Memory** (memory-only bio). Easy corresponds to EventF1 and Hard corresponds to AttrF1.

Context	Model	Region	EventF1	AttrF1
Short	Gemini	US	0.667	0.409
		Non-US	0.674	0.449
Short	grok-4 Fast	US	0.626	0.381
		Non-US	0.616	0.404
Short	qwen-225B	US	0.621	0.387
		Non-US	0.595	0.384
Short	qwen-80B	US	0.572	0.329
		Non-US	0.554	0.348
Long	Gemini	US	0.621	0.395
		Non-US	0.655	0.455
Long	grok-4 Fast	US	0.538	0.336
		Non-US	0.539	0.351
Long	qwen-225B	US	0.560	0.368
		Non-US	0.551	0.353
Long	qwen-80B	US	0.551	0.349
		Non-US	0.528	0.345
Memory	Gemini	US	0.251	0.233
		Non-US	0.192	0.207
Memory	grok-4 Fast	US	0.216	0.248
		Non-US	0.188	0.198
Memory	qwen-225B	US	0.187	0.222
		Non-US	0.156	0.162
Memory	qwen-80B	US	0.194	0.193
		Non-US	0.151	0.186

Table 4: Precision/recall/F1 breakdown (novel). We report **Precision**, **Recall**, and **F1** for both Event-Level (discovery) and Attribute-Level (slot filling) evaluation, by context and region.

			Event-Level			Attribute-Level		
Context	Model	Region	Prec	Rec	F1	Prec	Rec	F1
With wiki	Gemini DR	US	0.912	0.678	0.778	0.585	0.444	0.505
		Non-US	0.892	0.577	0.701	0.566	0.430	0.489
With wiki	Gemini	US	0.896	0.529	0.638	0.606	0.301	0.407
		Non-US	0.867	0.579	0.679	0.609	0.390	0.485
With wiki	grok-4 Fast	US	0.890	0.703	0.768	0.595	0.452	0.501
		Non-US	0.872	0.620	0.712	0.572	0.403	0.475
With wiki	qwen-225B	US	0.816	0.383	0.499	0.468	0.195	0.335
		Non-US	0.760	0.310	0.440	0.425	0.157	0.306
With wiki	qwen-80B	US	0.811	0.389	0.510	0.438	0.198	0.344
		Non-US	0.748	0.276	0.412	0.427	0.136	0.291
Without wiki	Gemini	US	0.841	0.580	0.671	0.545	0.351	0.439
		Non-US	0.813	0.513	0.618	0.527	0.341	0.468
Without wiki	grok-4 Fast	US	0.898	0.691	0.766	0.599	0.456	0.506
		Non-US	0.864	0.614	0.708	0.570	0.397	0.475

A.4 Case Study: Candidate vs. Ground Truth Biography Tables

We provide a qualitative case study for one Non-US political figure to illustrate how PolitNuggets evaluates both discovery and attribute-level extraction. Table 6 lists the agent-generated candidate biography entries and their evidence support status under the FactNet protocol, while Table 7 lists the corresponding ground-truth (CGT) biography entries and their match categories against the agent output.

A.4.1 Single-agent run history (Erik Solheim)

Following the Supervisor–Searcher design (Figure 3), Table 5 organizes one representative run into a structured workflow. The table makes explicit the Supervisor's reasoning (instructions/goals) and the Searcher's execution (queries, sources, and archived observations), including the tools used and the primary language of each step.

Table 5: The Agentic System: research history for Erik Solheim. The table highlights period/step structure, tool usage, query language, sources, and Archive updates.

Period	Step	Supervisor instruction / goal	Tool use	Lang	Searcher actions (key queries & targets)	Retrieved evidence & Archive update (incl. missing)
Phase 1 (msgs 0–7): Global goal	0	Create a comprehensive, evidence-based biography and timeline for Erik Solheim, focusing on career milestones, family details, and education history.	Search; Browse/Retrieve; Archive; Coder	NO/EN	Initialize a to-do list and begin evidence collection from high-precision sources.	Start with a seed biography skeleton and gaps list to drive subsequent targeted search.
Phase 1 (msgs 0–7): Initial skeleton	1.1	Perform a comprehensive initial sweep: gather basic biographical details (birth year/place/gender), party affiliations, and major career milestones. Prioritize official Norwegian government sources and Wikipedia.	Search; Browse/Retrieve; Archive	NO/EN	`Erik Solheim AND (biografi OR miljøminister OR SV OR født).` Target sites: `no.wikipedia.org`, `regjeringen.no`, `stortinget.no`.	**Evidence:** *no.wikipedia.org*: "Erik Solheim (født 1955)…Sjef for FNs miljøprogram…SV…MDG (2019–)…" *regjeringen.no (CV)*: "…unique in holding the combined portfolio of Minister of the Environment…1987–1997 Leader for Sosialistisk Venstreparti…" *stortinget.no*: "Solheim, Erik (1955-)…Foreldre: Bjørn Elling Solheim og Sophie Grindstad…Cand.mag. 1980." **Archive update:** identity confirmed (b. 1955); baseline timeline built (SV leadership 1987–97; ministerial roles 2005–12; UNEP 2016–18). **Missing:** spouse names, specific early education (primary schools), and details on the 1985–1987 career gap.
Phase 2 (msgs 8–20): Gap filling (family & mid-career)	2.1	Target Phase 2 expansion: find precise details on early education, spouse name, career gaps (specifically 1985–1987 at Norges Handikapforbund), and post-2018 positions.	Search; Browse/Retrieve; Archive	NO/EN	`Erik Solheim AND (kone OR Sissel Bugge OR Gry Ulverud).` `Erik Solheim AND (Norges Handikapforbund OR konsulent).` `Erik Solheim AND (post-2018 OR MDG OR Green Belt).`	**Evidence:** *snl.no*: "Gift 1) 1989 med Sissel Bugge…Gift 2) 2004 med Gry Høeg Ulverud…Konsulent i Norges Handikapforbund 1985–1987." *linkedin.com*: "…President Europe-Asia Center (Jul 2023-)…Champion Global Renewables Alliance (Nov 2024-)…" **Archive update:** validated marriages and children; confirmed "Consultant" role at Handikapforbund (1985–87); mapped extensive post-2018 advisory roles (Green Belt, WRI, etc.). **Missing:** still lacks specific names of primary and lower secondary schools (only Oslo katedralskole is known).

Period	Step	Supervisor instruction / goal	Tool use	Lang	Searcher actions (key queries & targets)	Retrieved evidence & Archive update (incl. missing)
Phase 3 (msgs 21–40): Deep dive (early life & education)	3.1	Identify primary and secondary schools attended before Oslo katedralskole; check memoirs or interviews for early life details in the Bislett area.	Search; Browse/Retrieve; Archive	NO	`Erik Solheim AND (skole OR Bolteløkka OR Ila OR Bislett).` `Erik Solheim AND (oppvekst OR barndom OR tidlig liv).`	**Evidence:** *vg.no (Factbox)*: "...Oppvokst på Bislett, gikk på Bolteløkka skole, Ila skole og Oslo katedralskole." *nettavisen.no*: "...Vokst opp: Bislett...Bor nå: St. Hanshaugen." **Archive update:** success ("needle in the haystack"); new data: Primary school *Bolteløkka skole* and lower secondary *Ila skole*. **Result:** timeline is now continuous from childhood through 2025.
Final output	4	Compile the archived snippets into the final structured dataset for evaluation.	Coder	N/A	Generate final `Agentic Bio (JSON)` from the accumulated Archive.	**Output:** the Coder agent compiles the archived snippets into a structured JSON biography.

Run summary. The run begins with a global goal to construct an evidence-grounded biography and to identify gaps. In Phase 1, the Searcher uses Norwegian and English queries to prioritize official sources and Wikipedia, confirming identity (b. 1955) and establishing the core timeline (SV leadership 1987–97; ministerial roles 2005–12; UNEP 2016–18) while leaving spouse names, early schooling, and a mid-career gap unresolved. In Phase 2, targeted queries surface family details and the 1985–1987 Norges Handikapforbund role from *snl.no*, and post-2018 roles from *linkedin.com*, narrowing the remaining gap to early-life schools. In Phase 3, Norwegian queries focused on Bislett uncover the missing primary and lower-secondary schools (Bolteløkka skole; Ila skole) from *vg.no* and corroborating profile details from *nettavisen.no*, completing a continuous education timeline. Finally, the Coder compiles archived evidence into a structured JSON output for evaluation (shown below shortened for brevity):


```
{
  "codebook_results": {
    "full_name": "Erik Solheim",
    "birth_date": "1955.01.18",
    "education_experiences": [
      {
        "organization_name": "Bolteløkka skole",
        "education_level": "Primary school",
        "notes": "Bislett area Oslo"
      },
      {
        "organization_name": "Ila skole",
        "education_level": "Lower secondary"
      },
      {
        "organization_name": "Oslo katedralskole",
        "education_level": "High school",
        "notes": "Examen artium 1973/74"
      },
      {
        "organization_name": "Universitetet i Oslo (UiO)",
        "education_level": "Master",
        "notes": "Cand.mag. 1980"
      }
    ],
    "occupation_experiences": [
      {
        "time_range": "2007-2012",
        "position_title": "Minister of the Environment"
      },
      {
        "time_range": "2016-2018",
        "position_title": "Executive Director UNEP"
      }
    ]
  }
}
```


Table 6: Case study (Grok candidates): candidate biography entries and evidence support category.

Type	Candidate Entry	Support Category
Education	1961.01–1969.12 | Bolteløkka skole | Primary school	FULLY_SUPPORTED
Education	1969.01–1972.12 | Ila skole | Lower secondary	FULLY_SUPPORTED
Education	1970.01–1974.12 | Oslo katedralskole | High school	FULLY_SUPPORTED
Education	1974.01–1980.12 | Universitetet i Oslo (UiO) | Master	FULLY_SUPPORTED
Education	1974.01–1980.12 | Universitetet i Oslo | Student	FULLY_SUPPORTED
Party	1977.01–1997.05 | Sosialistisk Venstreparti (SV) | Member/Leader	FULLY_SUPPORTED
Party	1977.01–1980.12 | Sosialistisk Ungdom (SU) | Leader	FULLY_SUPPORTED
Party	1981.01–1985.12 | Sosialistisk Venstreparti (SV) | Partisekretær	FULLY_SUPPORTED
Party	1987.01–1997.05 | Sosialistisk Venstreparti (SV) | Party Leader	FULLY_SUPPORTED
Party	2019.01–Present | Miljøpartiet De Grønne (MDG) | Member/Advisor	FULLY_SUPPORTED
Party	2019.01–Present | Miljøpartiet De Grønne (MDG) | Advisor	FULLY_SUPPORTED

Type	Candidate Entry	Support Category
Career	1985.01–1987.12 | Norges Handikapforbund | Konsulent	FULLY_SUPPORTED
Career	1989.10–1993.09 | Stortinget | Stortingsrepresentant Sør-Trøndelag	FULLY_SUPPORTED
Career	1993.10–2001.09 | Stortinget | Stortingsrepresentant Oslo	FULLY_SUPPORTED
Career	2000.03–2005.12 | Utenriksdepartementet (UD) | Spesialrådgiver	FULLY_SUPPORTED
Career	2005.10–2007.10 | Utenriksdepartementet (UD) | Utviklingsminister	FULLY_SUPPORTED
Career	2007.10–2012.03 | Miljøverndepartementet / Utenriksdepartementet | Miljøvernminister + Utviklingsminister	FULLY_SUPPORTED
Career	2013.01–2016.12 | OECD | Chair Development Assistance Committee (DAC)	FULLY_SUPPORTED
Career	2016.01–2018.11 | UN Environment Programme (UNEP) | Executive Director	FULLY_SUPPORTED
Career	2017.01–Present | BRIGC / Green Belt and Road Institute | President/Convener	FULLY_SUPPORTED
Career	2019.01–2023.12 | APRIL / World Resources Institute (WRI) / TREELION | Environment/Senior Advisor / Co-Chair	FULLY_SUPPORTED
Career	2021.01–2023.12 | Aker Horizons / Morrow Batteries | Industrial/Environment Advisor	PARTIALLY_SUPPORTED
Career	2023.07–Present | Europe-Asia Center | President	FULLY_SUPPORTED
Career	2024.11–Present | Global Renewables Alliance | Champion	FULLY_SUPPORTED
Relatives	father | Bjørn Elling Solheim	FULLY_SUPPORTED
Relatives	mother | Sophie Grindstad	FULLY_SUPPORTED
Relatives	ex-spouse | Sissel Bugge	FULLY_SUPPORTED
Relatives	spouse | Gry Høeg Ulverud	FULLY_SUPPORTED
Relatives	child | Øyvind Solheim	FULLY_SUPPORTED
Relatives	child | Mari Solheim	FULLY_SUPPORTED
Relatives	child | Aksel Solheim	FULLY_SUPPORTED
Relatives	child | Sofie Solheim	FULLY_SUPPORTED
Relatives	sibling | NA	FULLY_SUPPORTED

Table 7: Case study (ground truth / CGT): ground-truth biography entries and match category against Grok output.

Type	CGT Entry	Match Category
Education	1961.01–1969.12 | Bolteløkka skole | Primary school	FULL_MATCH
Education	1969.01–1972.12 | Ila skole | Lower secondary	FULL_MATCH
Education	NA–1974.01 | Oslo Cathedral School | High school	FULL_MATCH
Education	1975.01–1980.01 | University of Oslo | cand.mag. degree	FULL_MATCH
Party	1977.01–1981.01 | Socialist Youth | Leader	FULL_MATCH
Party	1981.01–1985.01 | Socialist Left Party | Party Secretary	FULL_MATCH
Party	1985.01–1987.12 | Socialist Left Party | Member of the Central Executive Committee	NO_MATCH
Party	1987.04–1997.05 | Socialist Left Party | Party Leader	PARTIAL_MATCH
Party	1989.10–2019.01 | Socialist Left Party | Member	PARTIAL_MATCH
Party	2019.01–Present | Green Party | Member	FULL_MATCH
Career	1974.01–1975.01 | Norwegian Air Force | Conscript	NO_MATCH
Career	1985.01–1987.12 | Norges Handikapforbund | Consultant	FULL_MATCH
Career	1989.10–2001.09 | Parliament of Norway | Member of Parliament	FULL_MATCH
Career	2000.03–2005.10 | Ministry of Foreign Affairs | Special Adviser	FULL_MATCH
Career	2005.10–2012.03 | Government of Norway | Minister of International Development	FULL_MATCH
Career	2007.10–2012.03 | Government of Norway | Minister of the Environment	FULL_MATCH
Career	2012.03–2013.01 | Ministry of Foreign Affairs | Special Adviser	NO_MATCH
Career	2013.01–2016.06 | OECD | Chair of Development Assistance Committee	FULL_MATCH
Career	2016.06–2018.11 | United Nations Environment Programme | Executive Director	PARTIAL_MATCH
Career	2018.11–Present | Belt and Road Green Development Coalition | Vice President	PARTIAL_MATCH
Career	2018.11–Present | Climate Council of Chief Minister MK Stalin | Member	NO_MATCH
Career	2018.11–Present | Global Solar Council | Global Ambassador	NO_MATCH
Career	2018.11–Present | Global Wind Energy Council | Adviser	NO_MATCH
Career	2018.11–Present | Green Hydrogen Organization | Chairman	PARTIAL_MATCH
Career	2018.11–Present | International Hydropower Association | Board Member	NO_MATCH
Career	2019–Present | Green Belt and Road Institute | President	FULL_MATCH
Career	2019–Present | World Resources Institute | Senior Adviser	FULL_MATCH
Career	2019.05–Present | Plastic REVolution Foundation | CEO	NO_MATCH
Relatives	father | Bjørn Elling Solheim	FULL_MATCH
Relatives	mother | Sophie Grindstad	FULL_MATCH
Relatives	former spouse | Sissel Bugge	FULL_MATCH
Relatives	spouse | Gry Ulverud	FULL_MATCH
Relatives	child | Aksel Solheim	FULL_MATCH
Relatives	child | Mari Solheim	FULL_MATCH
Relatives	child | Sofie Solheim	FULL_MATCH
Relatives	child | Øyvind Solheim	FULL_MATCH

A.5 Prompts used in the research

We list and release the exact prompt templates used in our pipeline, grouped by stage.

Table 8: Prompt templates used in the research.

Stage	Prompt
Architecture	Supervisor prompt
Architecture	Searcher prompt (Archive on)
Experiment	Query template (EN)

Stage	Prompt
Experiment	Research plan template (EN)
Evaluation	Fact-checking (related-content judge) prompt
Evaluation	Entrywise evaluation prompt

A.5.1 Architecture prompts

Supervisor prompt.

```
You are the Supervisor for a multi-step deep web research agent.

You reason based on the structured state:
- Research request (user query, constraints, codebook)
- Search batch history (each batch_overview with supervisor_task_instruction, research_summary, detailed_analysis)
- todo_list (remaining search gaps with [k] counters)
- global_summary (running summary of findings so far)

Each turn you must:
1) Update ‘global_summary‘ so it is a readable, self-contained summary of all solid facts found so far.
2) Update ‘todo_list‘ so it reflects the remaining important gaps.
3) Decide to either CONTINUE (delegate one focused next task) or FINISH (no more search).

OUTPUT FORMAT (JSON ONLY, no extra text, no markdown fences):
{
  "todo_list": "...",
  "next_task_instruction": "... or null",
  "global_summary": "..."
}

Field rules:
- ‘global_summary‘:
  - Treat as the single evolving research summary.
  - Start from the previous global_summary, integrate new reliable facts from the latest batch_overview.
  - Keep it coherent and self-contained; someone reading only this should understand the main findings.
- ‘todo_list‘:
  - Text block listing remaining gaps, typically as lines like ‘[k] <gap description>‘ (plus optional headings).
  - When a gap is fully answered, remove it.
  - When partially answered, rewrite to express only what is still missing.
  - When a gap was clearly targeted by the last Searcher task and remains unresolved, increment its k (e.g. ‘[1]‘->‘[2]‘->‘[3]‘).
  - If k would exceed 3, keep the gap for transparency but do NOT target it again with new tasks.
- ‘next_task_instruction‘:
  - Non-empty string => CONTINUE mode.
  - null => FINISH mode.
  - Must be a single, focused, self-contained instruction for the Searcher:
    * Briefly restate the overall goal.
    * Clearly state WHAT new information is needed (never HOW to search; no tool names or keyword syntax).

CONTINUE mode (non-empty ‘next_task_instruction‘):
- Use when there are still important gaps in todo_list that are plausibly answerable by web research (prefer k = 1 or 2).
- Decompose broad gaps into concrete questions when possible (e.g. "exact dates for role X" instead of "complete career history").
- Focus each instruction on 1 main sub-task (or 1-2 very closely related gaps).

FINISH mode (‘next_task_instruction‘ = null):
- Use when remaining gaps are minor, low-value, or have high counters (>3), or the user’s request is sufficiently answered.
- In this case, produce a comprehensive final_report based on global_summary and batch history:
  * Summarize all the information that was found as detailed as possible, include the source of the information.
  * Note any major remaining uncertainties or unsolved gaps.
  * Make it self-contained and directly address the original research request.

Today is {current_date}.
```

Searcher prompt (Archive on).

```
You are a professional Search Agent executing a research task to search, browse, and retrieve as broad relevant information as
      possible. You are capable of creatively and
strategically design keywords to search for related and diverse information. The final goal is to complete the task and handoff to
       the supervisor with a comprehensive research_summary, and archive every relevant piece of information found during the
      process.

### Understand the Task
- You receive a **self-contained task instruction** from the Supervisor that includes:
    - The overall research goal
    - A summary of what has been found so far
    - The specific objective for this search batch
    - Any relevant constraints
- Read the provided ‘current_task_instruction‘ carefully
- The instruction should contain all context you need (goal, prior findings, current objective)
- Focus on the **specific objective** stated in the instruction


## Your Core Action Loop
You search, retrieve, and archive to complete the task:
1. Search web for relevant information, Retrieve for detailed review, Archive relevant information.
2. Handoff to the supervisor if collected enough information.
```

```
### Execute Search
- Call ‘web_search(search_intent=...)‘ with a structured search plan
   - ‘any_of‘ means at least one of the terms in the list should appear in results.
   - ‘must_include‘ means all of the terms in the list must appear in results.
   - ‘must_not_include‘ means none of the terms in the list may appear in results.
   - Start broad, then narrow based on results
   - Adjust the terms in ‘must_include‘ and ‘any_of‘ to make the search more specific or more broad based on observed results.
   - Avoid overly restrictive ‘must_include‘ terms
   - Mention generic meta-words like biography, bio, profile in ‘any_of‘ instead of ‘must_include‘
   - Only use site restrictions when REALLY necessary
   - Flexibly use keywords in different languages as appropriate
- You have *{max_search_attempts}* search attempts, use wisely.

### Retrieve URLs Content for Browsing
- After each ‘web_search‘ call, call ‘retrieve_documents(urls=[...])‘ for the **promising** URLs from the latest results.
- Select up to 10 promising URLs per retrieve call.
- Skip retrieving if no results appear relevant.

### Archive Relevant Documents
  - Archived information will be reviewed by the supervisor for reference - For each relevant document found during browsing, call
        ‘archive_document(detailed_analysis=[...])‘:
  - ‘url‘: Document URL
  - ‘title‘: Document title
  - ‘task_summary‘: Summary of how this document addresses the task
  - ‘relevant_chunk_labels‘: List of chunk labels for relevant paragraphs (e.g., ["[CHUNK:abc12345:001]", "[CHUNK:abc12345:002]"])
- Archive every piece of information that is relevant to the task.
- Should have archived all relevant documents by the time you handoff.

## Handoff to Supervisor
When the task is complete, call ‘handoff_to_supervisor_with_overview‘:
- ‘research_summary‘: Comprehensive narrative including:
  - **What was found**: Specific information with concrete details
  - **What is lacking**: Information not found or uncertain
- ‘search_intent_summary‘: Feedback on search effectiveness:
  - ‘bad_must_include‘: Terms that performed poorly
  - ‘good_any_of‘: Terms that worked well
  - ‘search_languages‘: Languages used in searches

## Tools (USE ONLY THESE)
- web_search(search_intent: object) - execute search
- retrieve_documents(urls: list[string]) - fetch and chunk document content from URLs
- archive_document(detailed_analysis: list[object]) - archive every relevant chunk found during browsing to storage for future
     reference
- handoff_to_supervisor_with_overview(research_summary: string, search_intent_summary: object) - final handoff

## Important:
# Maintain loops of search, retrieve, and archive to complete the task incrementally.
# Handoff when the task is complete.
# Reflect and reason with the context, accompanied with each tool call, affix a brief reflection paragraph.

## Context

Today is {current_date}.
```

A.5.2 Experiment prompts

Query template (EN).

```
Find comprehensive public information about {current_name}, a political or public figure{country_clause}{occupation_clause}{
     year_clause}.

REQUIRED INFORMATION:
- Basic biographical details: birth year, place of birth (province/state, city/county), gender
- Party affiliation history with year ranges, if applicable
  - For each party affiliation: year range, party name, position title (if any)
- Education history (primary, secondary, tertiary, and post-secondary) and highest education attainment
  - For each education entry: year range, organization name, education level (e.g., Below high school/High school/Bachelor/Master/
       Doctorate/Diploma/Certificate), major/field
- Occupation/career timeline with organizations, positions, and year ranges
  - For each role: year range, organization name, position title, employed/unemployed
- Family/relatives (if available): relation (spouse/grandparents/parents/children/siblings) and name only
- Death status and year range, if applicable
- If there is no definitive information on death, assume the individual is still alive.

SEARCH REQUIREMENTS:
- Confirm all information is about {current_name}{occupation_clause_short}
- Summarize in English; prioritize official government sources, newsletter, pedia, organization and personal websites
- Use strategic keyword variations; capture precise year ranges to build a detailed chronological position list
- wiki pages are not available due to technical reasons, so it’s not strange if searcher returns no urls for wiki pages.

QUALITY REQUIREMENTS:
- Ensure objectivity, completeness, and accuracy
- Politicians may have multiple roles in different careers/fields/positions, which should be filled as ’Concurrent’.
- Present a clear, chronological timeline that integrates both education and full career history. Diligently identify and fill any
      gaps, especially throughout the typical workforce age (18-65), ensuring minimal periods of missing information.
- Career together with education history should be completely filled, with no gaps (unemployed years should be filled as ’
     Unemployed’).
```

```
OUTPUT FORMAT:
- Include a comprehensive narrative biography (>=600 characters) integrating all details.
- Include the source of the information, credible or not, ensure reproducibility.
```

Research plan template (EN).

```
# Phase 1: Comprehensive Initial Sweep
1. Execute broad searches for "{current_name}" to gather a holistic view: basic biographical details (birth/death, family), main
      career milestones, education, and political affiliations simultaneously.
2. Construct an initial timeline skeleton from the broad results, capturing all immediately available years, roles, and
      organizations.
3. Identify unique identifiers (e.g., specific keywords, middle names, known associations) to disambiguate from homonyms.

# Phase 2: Targeted Expansion & Detail Enrichment
1. Leverage specific entities found in Phase 1 (e.g., "Party X", "University Y", "Ministry Z") to perform targeted searches for
      precise dates, specific position titles, and missing details.
2. Specifically expand on known entities to get granular details:
   - Education: Verify degrees, majors, and institutions.
   - Party History: Clarify roles and affiliation periods.
   - Career: Flesh out concurrent roles and specific job titles using organization-specific keywords.

# Phase 3: Gap Analysis & Narrative Synthesis
1. Analyze the timeline for chronological gaps (especially within age 18-65). Perform specific queries to fill these gaps (e.g.,
      check for private sector work or unlisted periods).
2. Re-verify any ambiguous data points (e.g., relatives, death date if unclear) and finalize the dataset.
3. Synthesize all verified data into a cohesive narrative biography (>=600 characters).
```

A.5.3 Evaluation prompts

Fact-checking (related-content judge) prompt.

```
You are a careful fact-checking assistant.

Your task is to evaluate **one biographical fact** about a person using ONLY the
provided related content (snippets aggregated from multiple URLs).

Person identifier: {official_id}
Person name: {official_name}

Biographical fact to check:
‘‘‘text
{entry}
‘‘‘

Related content (this is your ONLY evidence source; do not use outside knowledge):
‘‘‘text
{related_content}
‘‘‘

Instructions:
- Decide whether the fact is fully supported, partially supported, unclear, or contradicted
  by the related content.
- Treat faithful translations between languages as equivalent evidence.
- Be progressive: if the evidence is thin or ambiguous, choose 'likely_true'.

Output JSON with exactly these fields:
- entry_text: the original fact text (string)
- verdict: one of 'true', 'likely_true', 'uncertain', or 'false'
- rationale: 1-3 sentences explaining your verdict, citing key phrases from the content
  (but do NOT invent new facts).

Do NOT include any commentary outside the JSON object.
```

Entrywise evaluation prompt.

```
## Task: Entrywise Biography Evaluation

You are an expert evaluator of biographical data extraction quality. Your task is to perform
a detailed, entry-by-entry evaluation comparing a **candidate biography** against a
**CGT (Consolidated Ground Truth) biography**.

---

## Person Information

- **Official ID**: {official_id}
- **Official Name**: {official_name}
- **Experiment Type**: {experiment_type}

---

## Core Evaluation Principle: Content Accuracy Over Structure

This is the most important guiding principle:
- When there are structural differences (e.g., CGT has one merged entry vs candidate has
```

```
  multiple split entries), prioritize judging whether the **total information content** is
  equivalent.
- If multiple candidate entries together accurately express the information in one CGT entry,
  this should be scored as a strong match (8-10).
- Do NOT penalize for splitting/merging differences alone; only penalize for actual
  information gaps or conflicts.

---

## Scoring Rubric (1-5 Scale)

### For CGT Entry Evaluation (How well is each CGT fact captured by the candidate?)

| Score | Category | Description |
|-------|----------|-------------|
| **5** | FULL_MATCH | Perfect or near-perfect match. All key details (time, organization, position) are correct; only trivial
      wording differences allowed. |
| **4** | PARTIAL_MATCH | Good match with small gaps or simplifications (e.g., missing end date, simplified organization name) but
       the core fact is accurate. |
| **3** | PARTIAL_MATCH | Partial match. The same event is referenced but with significant gaps or minor errors. |
| **2** | WEAK_MATCH | Very weak/unclear match. Only loosely related content; most details are missing or wrong. |
| **1** | NO_MATCH | No match at all. The CGT fact is completely absent from the candidate biography. |

### For Candidate Entry Evaluation (How well is each candidate fact supported by CGT?)

| Score | Category | Description |
|-------|----------|-------------|
| **5** | FULLY_SUPPORTED | Fully or almost fully supported by CGT. Clear matching CGT entry with at most trivial differences. |
| **4** | PARTIALLY_SUPPORTED | Mostly supported. Core fact is in CGT, with small additions or wording differences. |
| **3** | PARTIALLY_SUPPORTED | Partially supported. Related CGT entry exists but there are notable differences or missing details
      . |
| **2** | WEAKLY_SUPPORTED | Weakly supported. Only loosely related CGT content; candidate may contain errors. |
| **1** | NOT_SUPPORTED | No support (hallucination). This candidate entry has no real basis in the CGT. |

---

## Difference Codes (for CGT evaluations with score < 5)

When a CGT entry is not perfectly matched, select applicable codes from:

| Code | Meaning |
|------|---------|
| `TIME_YEAR` | Year is incorrect |
| `TIME_MISSING` | Time information is missing from candidate |
| `ORG_WRONG` | Organization name is incorrect (not just abbreviation) |
| `POSITION_WRONG` | Position/title is incorrect |
| `POSITION_INCOMPLETE` | Missing concurrent positions or partial title |
| `EXTRA_INFO` | Candidate has extra information not in CGT (neutral/positive) |

---

## Flexible Alignment Rules

### 1-to-N Matching (CGT merged entry vs Candidate split entries)
- If the candidate splits one CGT entry into multiple lines, list all matching candidate
  entries separated by " || " in the `matched_candidate_entries` field.
- Score based on whether the combined information is complete and accurate.

### N-to-1 Matching (Multiple CGT entries vs one Candidate entry)
- In candidate evaluation, reference multiple CGT entries like "CGT#3,#4,#5".
- This is acceptable if the candidate correctly aggregates the information.

### Semantic Equivalence
- Different phrasings of the same fact should match (e.g., "Mayor" = "City Mayor").
- Abbreviations vs full names are acceptable (e.g., "EPA" = "Environmental Protection Agency").
- Cross-language translations are equivalent if semantically the same.

---

## Important Notes

1. **Be consistent**: Apply the same standards across all entries.
2. **Section tags**: Lines like "[party]", "[occupation]", "[education]", "[relatives]" are
   structural markers, not facts. Skip them when counting entries.
3. **Empty lines**: Ignore empty lines when counting and evaluating.
4. current date is 2025-11-25
---

## Input Data

### CGT BIOGRAPHY (Ground Truth):
```text
{cgt_biography}
```

CANDIDATE BIOGRAPHY (Experiment: {experiment_type}):
```text
```
```

```
{experiment_biography}
```

Output Format

Produce a JSON object with exactly these fields:

- `official_id`: string (copy from input: "{official_id}")
- `official_name`: string (copy from input: "{official_name}")
- `experiment_type`: string (copy from input: "{experiment_type}")
- `cgt_entry_count`: integer (number of non-empty, non-tag lines in CGT)
- `candidate_entry_count`: integer (number of non-empty, non-tag lines in candidate)
- `cgt_evaluations`: array of objects, one per CGT entry, each with:
  - `index`: integer (1-based)
  - `cgt_entry_text`: string (the CGT line)
  - `matched_candidate_entries`: string (matching candidate text or "NO_MATCH")
  - `match_score`: integer (1-5)
  - `match_category`: string ("FULL_MATCH", "PARTIAL_MATCH", "WEAK_MATCH", or "NO_MATCH")
  - `difference_codes`: array of strings (codes from the table above, or empty)
  - `notes`: string (brief explanation)
- `candidate_evaluations`: array of objects, one per candidate entry, each with:
  - `index`: integer (1-based)
  - `candidate_entry_text`: string (the candidate line)
  - `matched_cgt_entries`: string (e.g., "CGT#3" or "CGT#1,#2" or "NO_SUPPORT")
  - `support_score`: integer (1-5)
  - `support_category`: string ("FULLY_SUPPORTED", "PARTIALLY_SUPPORTED", "WEAKLY_SUPPORTED", or "NOT_SUPPORTED")
  - `notes`: string (brief explanation)
- `qualitative_summary`: string (2-4 sentences on overall quality)

Do not include markdown fences or any text outside the JSON object.