Agentic Framework for Political Biography Extraction

Yifei Zhu* Songpo Yang† Jiangnan Zhu‡ Junyan Jiang§

Abstract

Producing large-scale political datasets demands extracting structured facts from unstructured sources, traditionally relying on expensive human experts and resisting at-scale automation. This paper develops and evaluates large language model (LLM)-based solutions to this bottleneck, focusing on elite biographies, one consequential class of political facts. We propose a two-stage "Synthesis–Coding" framework: LLM agents first search, filter, and curate evidence from heterogeneous web sources, then map curated inputs into structured records. We validate the framework across Chinese, American, and OECD political elites, benchmarking performance against human baselines using multiple state-of-the-art LLMs. We find that LLM coders match or exceed human experts when given curated inputs, and that agentic synthesis substantially outperforms human collective curation (Wikipedia) in open-web environments. We further identify a systematic bias: directly coding from long, multilingual corpora degrades extraction quality, and demonstrate that the synthesis stage mitigates this bias by compressing evidence into signal-dense representations.

*Ph.D candidate, Department of Politics and Public Administration, The University of Hong Kong. Email: frankyifei@connect.hku.hk.

†Boya Postdoctoral Fellow, School of International Studies, Peking University. Email: yangsp21@mails.tsinghua.edu.cn.

‡Associate Professor, Department of Politics and Public Administration, The University of Hong Kong. Email: zhujn@hku.hk.

§Assistant Professor, Department of Political Science, Columbia University. Email: jj3160@columbia.edu.

1 Introduction

The empirical revolution in political science has been fueled by the mass analysis of digitized political records (Gentzkow et al. 2019; Grimmer and Stewart 2013). Disclosed government documents, digitized news reports, and crawled web pages enable large-scale research using political facts[1], facilitating theory-building on representation, state capacity, and regime durability (Binderkrantz et al. 2024; Fisman et al. 2020; J. Jiang 2018; J. Jiang and M. Zhang 2020; Nyrup, Knutsen, et al. 2025). Yet transforming unstructured document stacks into structured, analyzable datasets remains prohibitively labor-intensive. The core bottleneck is fact extraction: gathering evidence from sources, extracting specific information, and structuring verifiable information into data suitable for downstream analysis requires extensive trained manual labor. Scaling political data production beyond this bottleneck demands automated solutions to replicate, or exceed, the validity of human coding while slashing labor costs.

This paper develops and evaluates large language model (LLM)-based solutions for extracting political facts from unstructured documents at scale, focusing on one consequential class of political facts: structured elite biography. Elites, shaped by different backgrounds, incentives, and networks, systematically influence policy-making, public opinion, and political stability across regime types (Alexiadou 2015; J. Jiang 2018; King et al. 2013; Putnam 1976; Reuter and Szakonyi 2019; Svolik 2012; Woldense and Kroeger 2024). Structured elite biographies are analytically rich, but computationally prohibitive using traditional manual methods. Manually constructing political biographies requires gathering and identifying appropriate sources and extracting biographical facts, including entities, events, and relations, from unstructured text into temporally organized lists.[2] In a landmark study, Nyrup, Knutsen, et al. (2025) mobilized over 30 research assistants (RAs) across three years to manually assemble the "Paths to Power" (PtP) dataset on cabinet members

[1] By "political facts," we refer to verifiable, descriptive attributes of political actors and institutions, including office-holding, educational backgrounds, career paths, and institutional affiliations, that can be reliably documented and cross-checked across sources.

[2] While we focus on elite biographies, the challenges identified here, such as information dispersion, temporal inconsistency, source conflict, and extraction from unstructured text, apply to other "narrative" political data, such as tracking policy evolution, coding event data from news reports, or reconstructing negotiation processes from diplomatic cables.

worldwide from diverse web resources.[3] While these achievements, together with many similar efforts (Armstrong et al. 2024; Bäck et al. 2021; J. Jiang 2018; Lee and McClean 2022; Raleigh and Wigmore-Shepherd 2022; Vittori et al. 2023), are indispensable for answering theoretical questions across diverse political contexts, the need for manual extraction imposes severe limitations, such as discontinued or out-of-date datasets, costly coverage expansion, and new-variable addition.

LLMs offer a potential path toward scalable political-text processing (Benoit, De Marchi, et al. 2025; Gilardi et al. 2023; Ornstein et al. 2025; Palmer et al. 2024). However, existing political-science applications focus on *classification* tasks, where the output is one label from a predefined finite set (Benoit, De Marchi, et al. 2025; Halterman and Keith 2024; Ziems et al. 2024). The harder question is whether LLM-based systems can perform valid extraction tasks, with multiple field codebooks and undefined output spaces. Extraction in political science also features long, noisy document collections and even web resources, which naive zero-shot or few-shot fails to address (N. F. Liu et al. 2024).

To automate extraction without compromising validity, we propose and evaluate an agentic framework for political-fact extraction from web sources. The framework comprises two stages: an upstream *synthesis* stage that utilizes agentic recursive LLM calls to search and refine evidence from web sources, and a downstream *coding* stage that maps that refined evidence into structured facts. We evaluate this framework against human extraction using a validated ground-truth dataset of political-elite biographies from China, the United States, and a comparative sample of OECD countries. We first validate the *coding* ability of LLMs when the input is human-curated Wikipedia biographies, showing that LLM coders can match and exceed human coding quality using curated short corpora. We then prove that the agentic workflow outperforms human collective synthesis (Wikipedia) in producing curated biography corpora for global political elites. Finally, we diagnose a systematic bias: long, multilingual corpora degrade extraction quality. By holding the retrieved evidence fixed and varying only its representation, we demonstrate that the synthesis

[3]The scale of such efforts is often understated. Coding a comprehensive cabinet dataset typically requires tens of thousands of research-assistant hours. For instance, the WhoGov dataset (Nyrup and Bramwell 2020) took nearly a decade of intermittent work.

stage mitigates this degradation by compressing evidence into signal-dense inputs.

Our contribution is fourfold. First, we formalize political-fact extraction as distinct from classification. Whereas classification typically assigns texts to a fixed set of labels, extraction requires identifying and structuring facts from open-domain sources and assembling them into temporally ordered structured records. Second, we propose a Synthesis-then-Coding framework for extraction, which treats information synthesis as a prerequisite for valid extraction and clarifies why skipping synthesis (e.g., naive context-window stuffing) induces a quantity–quality trade-off that degrades performance. Third, we develop and open-source a scalable agentic package that operationalizes synthesis through iterative, tool-using retrieval and refinement, and we show that it can outperform human collective synthesis (Wikipedia) in producing high-signal corpora for global political elites. Fourth, we apply this framework to generate a large cross-national dataset of political-elite biographies, lowering the barrier to producing and maintaining high-quality data in information-poor environments and providing a generalizable template for extracting structured narratives from unstructured text. Finally, our evaluation method based on verifiable political facts and synthetic ground truth provides template for evaluating future LLM agent application.

2 The Challenge of Extracting Political Biographies at Scale

To understand why automatic solutions are necessary, it is useful to discuss the type of political facts that are most challenging for large-scale data production. Among the many types, elite biographical information (e.g., who political elites are, where they come from, and how they advance through institutions) constitutes a particularly demanding case. Such data form the backbone of comparative political research. Granular information on educational backgrounds, career trajectories, and kinship networks is central to theories of political representation (Carnes 2024; Lee and McClean 2022), authoritarian power-sharing (Raleigh and Wigmore-Shepherd 2022; Svolik 2012), and technocratic governance (Lin 2020; Vittori et al. 2023). For example, detailed career histories have enabled Chinese political scholars to uncover the logic of factional patronage (J. Jiang 2018;

Shih et al. 2012) and to assess regime claims of meritocratic selection (H. Liu 2024).[4]

Despite their importance, elite biographical data remain exceptionally difficult to produce at scale. While digitization has expanded access to political texts,[5] extraction still relies overwhelmingly on manual coding by experts or RAs. Crowdsourcing platforms offer an alternative labor model, but are generally unsuitable for complex elite-data extraction, which requires substantial domain knowledge to resolve ambiguities in names, titles, and political affiliations (Benoit, Conway, et al. 2016). These constraints translate into extraordinary labor requirements. The PtP dataset (Nyrup, Knutsen, et al. 2025), covering cabinet ministers in 141 countries over 55 years, required more than five years of coordinated work by over 30 RAs. The LEAD dataset (Ellis et al. 2015) assigned multiple coders to each leader and still took three years to complete. Other prominent efforts, including Funke et al. (2023) and Braun and Raddatz (2010), likewise required years of intensive manual verification. As summarized in Table 1, high-quality elite datasets typically mobilize large teams over extended periods and depend on sustained funding from agencies such as the European Research Council or US National Science Foundation (Alexiadou 2022; Ellis et al. 2015).

Even with these investments, manual production has systematic limitations. First, intercoder reliability remains imperfect. For example, Nyrup, Knutsen, et al. (2025) report intercoder reliability of around 0.80 for cabinet-level biographical attributes, falling below 0.70 for certain variables.[6] More importantly, most datasets remain static snapshots. Among the datasets surveyed, only a small fraction have been updated in the past five years, while widely used resources such as Archigos (Goemans et al. 2009) and LEAD (Ellis et al. 2015) have remained unchanged for a decade or

[4]While our discussion focuses on national-level elites (e.g., cabinet ministers) due to data availability, the theoretical importance of biographical data extends to local officials, bureaucrats, and party cadres (Landry 2008). The scalability constraints we identify are arguably even more severe for these lower-tier populations, where N is larger and data are noisier.

[5]We distinguish between *digitization* (converting physical records into digital text) and *extraction* (converting unstructured text into structured databases). The latter remains the primary bottleneck for narrative political data.

[6]Similar levels of coder disagreement are reported in other elite datasets, including LEAD (Ellis et al. 2015), Archigos (Goemans et al. 2009), and WhoGov (Nyrup and Bramwell 2020), where resolving inconsistencies often requires multiple coding rounds or adjudication by senior researchers. More broadly, methodological surveys of text and manual data construction highlight that human hand-coding can introduce measurable errors and biases, motivating audits, cross-validation, and supervised approaches (Gentzkow et al. 2019; Grimmer and Stewart 2013).

Table 1: Available Datasets on Political Elites Since 2007*

Dataset	*N* Countries	Years†	Region	Variables (D/C/I/P)	Production Details‡
National Leaders					
Baturo (2016)	–	1960–2010	Global	✓/✓/–/✓	2 RAs, 2009–14
Baturo and Elkink (2022)	–	1950–2017	Global	–/✓/–/–	1 RA
Baturo and Tolstrup (2023)	132	1918–2019	Global	–/–/–/✓	Built on 11 datasets
Bomprezzi et al. (2025)	177	1989–2018	Global	✓/–/–/✓	27 RAs, 1.6M entities
De Luca et al. (2018)	140	1992–2013	Global	✓/–/–/–	Augmented Archigos
Dreher et al. (2009)	72	1970–2002	Global	✓/✓/–/–	–
Ellis et al. (2015)	188	1875–2004	Global	✓/✓/–/–	2 RAs/leader, 3 yrs
Eschenauer-Engler and Herre (2023)	–	1950–2020	Global	–/–/–/✓	–
Fearon et al. (2007)	161	1945–1999	Global	✓/–/–/–	–
Funke et al. (2023)	60	1900–2020	Global	–/–/✓/–	9 RAs, 20k+ pages
Gerring et al. (2019)	162	2010–2013	Global	✓/✓/–/–	–
Goemans et al. (2009)	188	1875–2015	Global	–/–/–/✓	–
Herre (2023)	182	1945–2020	Global	–/–/✓/–	15 RAs
Licht (2022)	–	1960–2015	Global	–/–/–/✓	5 grad students
Mattes et al. (2016)	169	1919–2018	Global	–/–/–/✓	9 RAs+Experts
Yu and Jong-A-Pin (2020)	177	1946–2011	Global	✓/✓/–/–	–
Sub-National & Ministerial-Level Elites					
Alexiadou (2015)	18	1945–2013	OECD	–/✓/–/–	–
Alexiadou (2022)	18	1945–2015	OECD	✓/✓/–/–	Multiple coders, 6 yrs
Alexiadou et al. (2022)	13	1980–2014	W. Europe	–/✓/–/–	Multiple experts
Armstrong et al. (2024)	191	1972–2017	Global	✓/✓/–/–	6 RAs
Bäck et al. (2021)	13	1789–2021	Great Powers	✓/✓/–/–	–
Braun and Raddatz (2010)	154	1996–2005	Global	–/–/–/–	72,769 names checked
Carozzi and Repetto (2016)	1	1994–2006	Italy	✓/–/–/–	–
Fuchs and Richert (2018)	23	1967–2012	OECD	✓/✓/✓/–	10 RAs
Hallerberg and Wehner (2012)	27	1973–2010	OECD	✓/–/–/–	6 RAs
J. Jiang (2018)	1	1997–2015	China	✓/✓/–/✓	20+ RAs
Lee and McClean (2022)	4	1983–2017	Asia	✓/✓/–/–	–
Nyrup and Bramwell (2020)	177	1966–2023	Global	–/–/✓/–	9+ coders
Nyrup, Knutsen, et al. (2025)	141	1966–2021	Global	✓/✓/–/–	30+ RAs, multi-year
Raleigh and Wigmore-Shepherd (2022)	23	1996–2017	Africa	–/–/–/–	–
Ricart-Huguet (2021)	16	1960–2010	Africa	✓/–/–/–	–
Vittori et al. (2023)	31	2000–2020	EU+4	✓/✓/–/–	Country experts

*This list is illustrative not exhaustive. We prioritize datasets that (1) focus on individual-level attributes of political elites, (2) are widely cited in top political-science journals, and (3) involve substantial manual coding efforts. Datasets focused solely on voting records (e.g., roll-call data) are excluded as they represent a different class of "atomic" facts.

†*Years* reflects the temporal coverage of the most recent available version. The end year indicates when the dataset was last updated, which may postdate the cited foundational paper (e.g., Archigos 4.1, initially published by Goemans et al. (2009), was subsequently updated to cover leaders through 2015). Most datasets have not been updated for several years, with many remaining frozen a decade or more behind current events.

‡*Variables*: D = Demographics (education, ethnicity, birthplace, family background); C = Career (pre-office occupation/political experience); I = Ideology/party affiliation; P = Power dynamics (entry/exit manner, tenure, transitions). For an illustration of what these dimensions look like as structured biography entries, see Table 2 in Section 4. Production scale indicates the reported labor intensity of manual data collection; – = Not reported or insufficient detail.

more. Updating comprehensive elite datasets often requires thousands of additional labor hours, making continuous maintenance prohibitively costly. These production constraints shape the substantive scope of political inquiry. Existing datasets disproportionately focus on actors at the apex of political power, while mid-level bureaucrats, local officials, and other actors central to policy implementation remain largely absent from comparative data. Even among top-tier elites, coverage is uneven: while finance ministers (Armstrong et al. 2024) and foreign ministers (Bäck et al. 2021) are relatively well documented, systematic data on portfolios such as education, health, or infrastructure remain scarce.

The structure of available information further constrains what can be collected. When comprehensive Wikipedia biographies exist, researchers can extract structured facts from consolidated text. Such cases, however, are unevenly distributed across countries and government levels and often omit early careers, family ties, or post-tenure activities. In their absence, researchers must reconstruct careers by manually searching government websites, news archives, and organizational announcements. Information is then fragmented across heterogeneous sources, frequently in local languages and embedded in noise. Hence, empirical research tends to concentrate where information is most accessible rather than where theoretical questions are most consequential, producing an "information structure bias" (Wilson and Knutsen 2022).

These realities expose two fundamental bottlenecks for any automated solution. The first is cost and scalability: manual coding scales linearly with dataset size, limiting expansion beyond narrow elite populations and hindering timely updates. The second is transparency and replicability: manually assembled datasets typically release only final records, with limited documentation of sources, conflicts, or adjudication rules, complicating verification and reuse. These constraints restrict not only the *scale* of political data production but also its *verifiability*, a growing concern as comparative political science increasingly relies on large-N observational evidence.

These challenges extend beyond elite studies. Scholars studying contentious politics increasingly rely on real-time event data-scraping from news sites and social media (King et al. 2013; Muthiah et al. 2015; H. Zhang and Pan 2019). Legal and regulatory research requires tracking

policy evolution across fragmented official gazettes, court databases, and agency announcements (Baturo, Dasandi, et al. 2017; Fang et al. 2025; Liebman et al. 2020). Economic research depends on synthesizing information from corporate filings, diplomatic cables, and industry publications (Hassan et al. 2019; Thrall 2025). The common core challenge is transforming vast, unstructured, often conflicting information into valid, structured datasets at scales that manual coding cannot sustain. Political-fact extraction, therefore, represents less a niche technical problem than a fundamental bottleneck constraining the empirical scope of comparative political science.

3 From Classification to Extraction: The Context Challenge

If manual coding is the bottleneck, advances in generative language models offer a theoretical solution. A growing body of work demonstrates that LLMs can reliably replicate human judgment on classification tasks such as ideology scaling (Wu et al. 2023), stance detection (Benoit, De Marchi, et al. 2025; Gilardi et al. 2023), and topic classification (Ornstein et al. 2025).[7] These classification tasks share a common structure: they map bounded, pre-selected texts into finite label sets, holding the input document fixed. The model receives a well-defined text—a speech, a manifesto, a social media post—and must interpret its content according to a predefined codebook.

Political-data extraction, however, represents a fundamentally different computational problem.[8] Unlike classification, which assigns labels to fixed texts, extraction entails actively seeking and reconstructing structured facts from vast, dispersed information environments where no single document contains complete information. This shift introduces four compounding challenges that classification benchmarks do not address.[9] First, extraction is *open-domain*: relevant entities,

[7]Empirically, LLMs not only match but often exceed humans on political-text classification. Gilardi et al. (2023) find that ChatGPT's zero-shot accuracy surpasses crowdworkers by approximately 25 percentage points on tasks involving stance, topics, and frame detection, while achieving higher intercoder agreement than trained human coders. Benoit, De Marchi, et al. (2025) show that LLM ratings of party manifestos correlate with expert judgments at 0.87–0.92, reaching the upper bound of human expert agreement, with intra-LLM consistency typically exceeding 0.90 versus human intercoder reliability of 0.3–0.5. Likewise, Wu et al. (2023) demonstrate that LLM-generated ideology scores achieve test–retest correlations of 0.997 and better predict human perceptions of politician ideology than traditional behavioral measures. These studies establish that modern LLMs already deliver both superior accuracy and consistency on classification tasks.

[8]We provide a formal mathematical distinction between classification and extraction in Online Appendix A1.

[9]In the natural language processing literature, this class of problems is commonly termed *open-domain slot filling*

organizations, and position titles are not exhaustively enumerated ex ante, so the system must recognize and standardize an effectively unbounded set of possible answers rather than selecting from a fixed menu. Second, extraction exhibits high *task complexity*: a single record (one official's career) requires answering many heterogeneous sub-questions (e.g., identity resolution, appointment dates, organizational affiliations, position titles, status flags), and reconciling contradictions across sources, rather than producing a single label. Third, extraction involves *context dependency*: information retrieval is inherently path-dependent. Discovering one fact (e.g., an official served as "Assistant Secretary at Commerce") provides the contextual cue necessary to locate subsequent facts (e.g., searching for "Commerce Assistant Secretary 2015" rather than the initial broad query "John Smith government").[10] Unlike independent classification labels, career events form temporal sequences where reconstructing one fact often depends on establishing another first.

These three challenges are formidable, and current evidence does not establish that LLMs can reliably address them at scale in real-world extraction scenarios. Theoretically, the same generative capabilities that enable classification could extend to these problems, yet translating this theoretical potential into validated extraction systems remains an open question. Even perfect solutions to the first three challenges would not resolve a fourth constraint that is orthogonal to model capability: the *long-context problem*. In open-ended environments such as the open web, potentially relevant information for a single individual is dispersed across hundreds or thousands of documents (far exceeding the token budgets of even extended-context models), and is submerged in vast quantities of irrelevant content. This challenge operates on two dimensions, neither of which can be resolved through straightforward technical improvements.

First, models may fail to locate related information or generate false information to reconcile contradictory or misleading signals in over-long contexts (N. F. Liu et al. 2024; Mallen et al.

or *complex information extraction* (Angeli et al. 2015). Unlike classification tasks with predefined label sets, these approaches aim to recover canonical attribute values from unstructured text under a specified codebook.

[10]This dependency mirrors the challenge of *multi-hop question answering* (Yang et al. 2018), where answering a query requires aggregating evidence from disjoint text segments. For biographical data, a resignation date in one document may resolve the tenure end date for a position mentioned in another, but only if both documents are located and linked.

2023; F. Shi et al. 2023).[11] Second, and more fundamentally, there exists a *capacity constraint* that no architectural refinement can overcome: even models with 100k+ token windows cannot accommodate the full universe of potentially relevant documents for a given extraction target. A mid-level bureaucrat in a large country may be mentioned across thousands of government websites, news articles, and policy documents spanning decades; web searches routinely return result sets that, if naively concatenated, would exceed 1 million tokens. These findings reveal that the binding constraint in open-ended extraction is not merely reading long text, but deciding which sources to read and how to condense them into high-signal inputs before structured coding. Valid extraction at scale therefore requires an architectural solution that governs information selection and condensation *before* LLM-based coding can be meaningfully applied.

4 An Agentic Solution to the Extraction Challenge

Building on the preceding analysis, we introduce an agentic architecture that directly targets the upstream bottleneck of evidence acquisition. The core idea is to decompose extraction into two analytically distinct stages: *synthesis*, which locates, evaluates, and consolidates relevant evidence from open-ended sources, and *coding*, which extracts structured facts from curated inputs. We operationalize this design through a recursive retrieval-and-synthesis loop that mirrors the iterative logic of human research and enables valid extraction from noisy web environments. Before describing the architecture in detail, Table 2 illustrates what the target output of this pipeline looks like: a structured biography decomposed into the DCIP dimensions introduced in Table 1, represented as a sequence of timestamped entries of the form `start-end | entity | role`. The example draws on Erik Solheim, a Norwegian minister whose biography is absent from existing

[11] N. F. Liu et al. (2024) demonstrate a U-shaped performance curve in multi-document question answering: accuracy peaks when relevant information appears at the beginning or end of long contexts but degrades sharply (often by over 20 percentage points) when the same information is positioned in the middle. This "lost-in-the-middle" phenomenon persists across model families (GPT-3.5, Claude, open-source alternatives) and is not resolved by simply extending context windows: models with 16k-token capacity perform identically to their 4k counterparts when processing inputs that fit in both, indicating that raw capacity does not translate into robust information use (N. F. Liu et al. 2024, p. 162).

hand-coded datasets and was recovered by the agentic framework described below.

Table 2: The DCIP Dimensions of Structured Elite Biography

DCIP	Dimension	Example structured biography entries
D	Demographics	*Education*: 1975.01–1980.01 | University of Oslo | cand.mag. *Relatives*: spouse; child
C	Career	1989.10–2001.09 | Parliament of Norway | Member of Parliament 2005.10–2012.03 | Government of Norway | Minister of International Development
I	Ideology/party affiliation	1981.01–1985.01 | Socialist Left Party | Party Secretary 2019.01–Present | Green Party | Member
P	Power dynamics	*Tenure/portfolios*: 2005.10–2012.03 | Cabinet minister | International Development 2007.10–2012.03 | Cabinet minister | Environment

Notes. Each dimension is operationalized through one or more biography entries of the form `start-end | entity | role`. Concurrent appointments (e.g., holding two cabinet portfolios simultaneously, as in the **P** row) are represented as overlapping entries sharing the same time range. The examples are drawn from Table A5.3, which reports the agentic synthesis output for Erik Solheim, former Minister of International Development in Norway—an official absent from existing hand-coded datasets and recovered by the framework introduced in this section.

Producing a record of this form for an official like Solheim—whose career spans parliamentary roles, cabinet portfolios, and post-tenure international appointments across multiple languages and source types—is precisely the extraction challenge this section addresses. The central difficulty lies in deciding *which* sources to consult and *how* to compress dispersed evidence into inputs suitable for downstream coding. Standard retrieval-augmented generation (RAG) systems address this by retrieving document chunks based on semantic similarity to a query, then passing concatenated results to the model (P. Lewis et al. 2020). In web-search scenarios, this amounts to issuing a single-keyword query, retrieving the top-ranked pages, and extracting the aggregated results. This one-shot approach is fundamentally brittle under the above-identified *context dependency* challenge. A document's relevance often is unknown ex ante but depends on information uncovered in prior retrieval steps. Key entities, affiliations, and career transitions are frequently discoverable only after intermediate facts have been established, rendering fixed retrieval strategies systematically incomplete.

Consider the biographical reconstruction task in Figure 1. When Wikipedia contains a comprehensive biography (left panel, green boxes), a single LLM pass suffices to extract structured facts. However, when Wikipedia is absent or incomplete (the common case for non-elite officials),

the system must search across heterogeneous web sources (right panel, blue boxes). Crucially, informative follow-up queries are endogenous to what has been learned from earlier documents. Discovering that an official served as "UNEP Executive Director" provides the contextual anchor needed to locate subsequent positions ("Climate Council member"), professional affiliations ("Belt & Road Coalition Vice-President"), or organizational roles ("Plastic REV Foundation CEO") that would be invisible to an initial broad search. Static RAG, by committing to a fixed retrieval strategy before any evidence has been examined, cannot exploit these path-dependent cues.

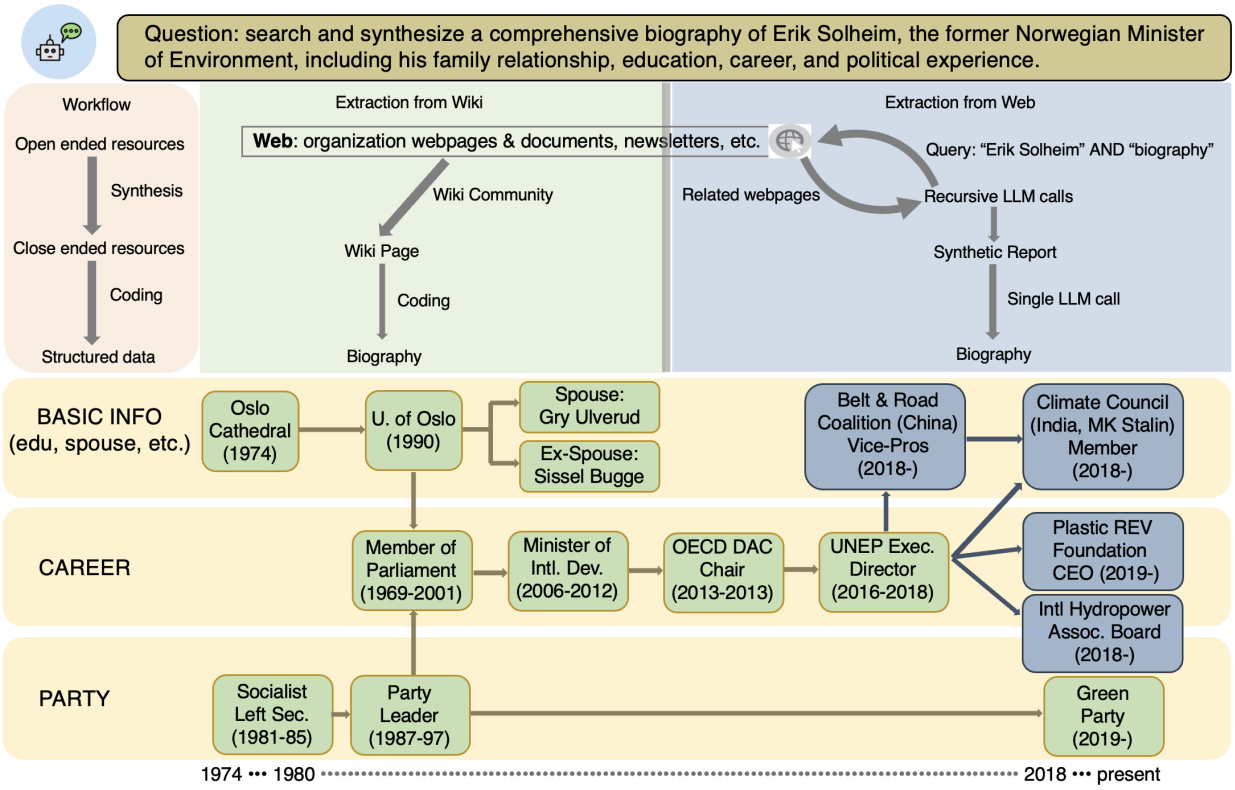

Figure 1: Two coding strategies for elite biographies. Left: when a Wikipedia page exists, we code directly from the curated page with a single LLM pass. Right: when Wikipedia information is missing or incomplete, we search across web sources and iteratively synthesize a synthetic report, then code from that report. The lower panel illustrates the structured output as an ordered biography (career, education, and affiliations) anchored on a timeline. This contrast highlights why extraction from open-web sources requires adaptive synthesis rather than one-shot retrieval.

RAG emphasizes the model's capacity to *observe*: given curated context, LLMs can reliably produce structured outputs in zero-shot or few-shot settings (Benoit, De Marchi, et al. 2025; Gilardi et al. 2023; Ornstein et al. 2025). However, modern LLMs can also *act*. Specifically, they can

generate executable commands that interact with external retrieval systems, enabling autonomous information-gathering. In recursive settings, the interleaving of action and observation allows the model to search iteratively, examine retrieved documents, reason about gaps in current knowledge, and decide what sources to consult next (Yao et al. 2023). Each retrieval action is thus conditioned on information accumulated in previous steps, allowing the system to resolve the path-dependent nature of open-domain retrieval while progressively compressing a noisy information universe into a condensed corpus that avoids long-context constraints.

We operationalize this capability through an *agentic framework* that repositions the LLM from passive reader to active assistant.[12] Rather than treating retrieval as preprocessing, the architecture implements a recursive reasoning–action loop: it iteratively (i) *reasons* about current knowledge gaps (e.g., "I have identified the UNEP directorship but lack information on prior ministerial roles"), (ii) *acts* to acquire missing evidence via targeted search queries or document inspection, and (iii) updates a running *synthetic report* consolidating verified findings. Each step is executed through a minimal set of deterministic retrieval tools, which carry out machine-readable commands (e.g., `search("Erik Solheim OECD DAC")` or *open_url*(*url*)) and return text for inspection. The agent iteratively incorporates evidence, decides whether further retrieval is needed, and maintains only the task description, recent interaction history, and current report in context to ground search decisions while avoiding context overflow. The final synthetic report, a compressed, wiki-like summary of curated evidence, serves as the sole input to the downstream coding step that produces the structured biography. The exact prompt templates used for the supervisor, searcher, and coder agents are listed in Online Appendix A6.

The agentic framework combines scalability, transparency, and validity in automated data production.[13] It can process thousands of targets in parallel without task-specific model training or human supervision, substantially reducing time and labor costs. Moreover, every retrieved source is archived and linked to the generated synthetic report, allowing researchers to inspect intermediate

[12]Full implementation details, including model specifications, operational constraints, search budgets, and cost breakdowns, are documented in Online Appendix A2.

[13]For a practical guide to applying this framework to new extraction tasks, including step-by-step recommendations on codebook design, synthesis configuration, and evaluation, see Online Appendix A7.

evidence, trace how claims were verified or adjudicated, and identify potential errors or biases (Bail 2024).[14] To evaluate whether these architectural advantages translate into valid and scalable data production, we design a multi-stage empirical evaluation. Figure 2 summarizes the agentic system architecture and the experimental contrasts. Experiment 1 tests whether LLMs can accurately code structured facts from curated biographical texts. Experiment 2 examines whether agentic synthesis from open-web sources can recover reliable biographical information in fragmented, noisy information environments. Experiment 3 then assesses the architectural mechanisms underlying these results by holding the retrieved evidence fixed and varying how that evidence is represented to the coder. This diagnostic experiment clarifies why synthesis is essential for reliable extraction at scale.

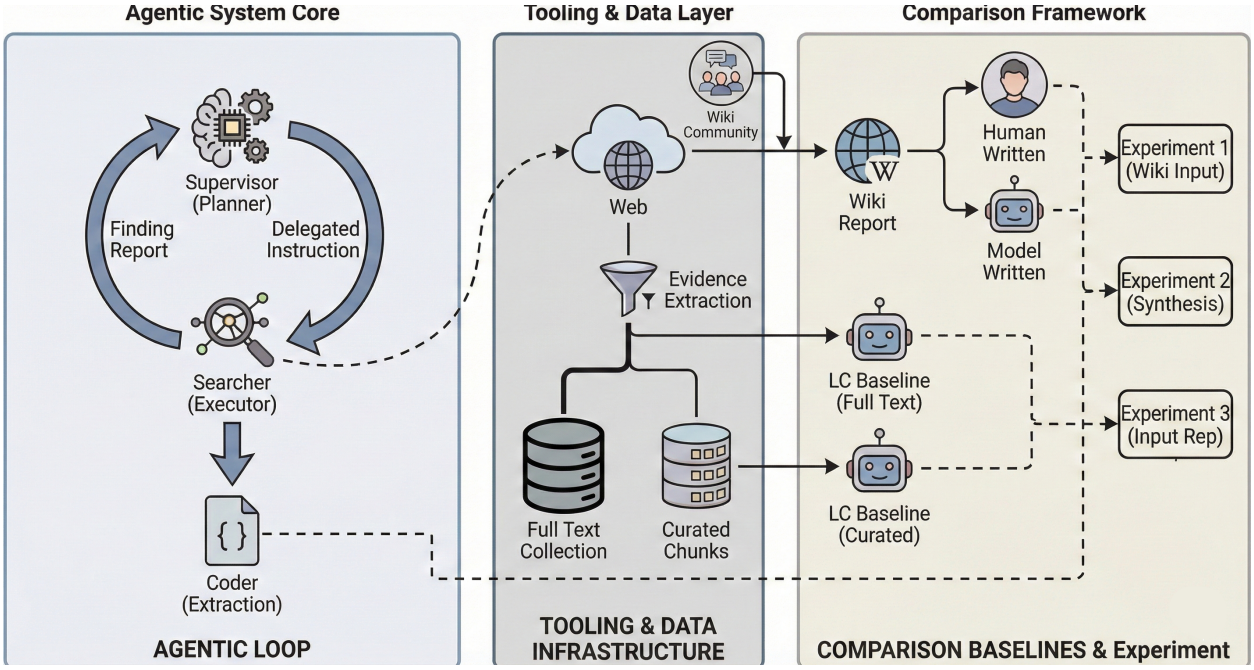

Figure 2: Architecture for agentic synthesis and experimental design. Left: the Supervisor–Searcher–Coder loop, in which a Supervisor maintains global state and delegates bounded retrieval tasks to specialized Searcher workers; retrieved evidence is stored in an Archive and mapped by a Coder into a structured biography. Right: the three experimental comparisons.

[14]For a complete step-by-step trace of an agentic extraction run for a specific official, see the case study of Erik Solheim in Online Appendix A5.

5 Experiment 1: The Coding Challenge

This section addresses the first research question by isolating the *coding* component of automated biography extraction. The objective is to assess whether, given identical and curated evidence, LLMs can extract structured biographical facts with accuracy comparable to human coders. By holding the information environment constant and varying only the coder, this design directly tests whether coding itself constitutes a binding constraint in automated political-data production. We implement this test in a setting where authoritative human-coded benchmarks exist, enabling direct validation of event-level extraction accuracy.

5.1 Data

Human-Coded Benchmark: The CPED The Chinese case provides a uniquely suitable benchmark for evaluating coding validity because it offers high-quality, human-coded biographical data at scale. Our analysis relies on the Chinese Political Elite Database (CPED), a comprehensive biographical database covering more than 4,000 key city-, provincial-, and national-level leaders since the late 1990s (J. Jiang 2018). Outside the Central Organization Department's internal archives of the Chinese Communist Party, the CPED is widely recognized as the most authoritative digital repository of Chinese political curricula vitae.[15] Importantly, all biographies in the CPED are manually coded by trained RAs following standardized rules, producing structured career histories that serve as a ground-truth benchmark for validation.

Sample Construction From the CPED population, we construct a stratified random sample of 197 officials, balanced across three administrative ranks to capture variation in career complexity: (i) bureau-director level (equivalent to city mayors or provincial department heads), (ii) vice-ministerial level (provincial governors or vice-ministers), and (iii) ministerial level (provincial party secretaries or national ministers). This stratification ensures representation across the hi-

[15]The CPED provides detailed information on career trajectories, educational backgrounds, native place, birth year, ethnicity, and records of corruption investigations.

erarchy of Chinese bureaucratic advancement, where career paths differ systematically by rank. These officials exhibit complex, longitudinal career histories typical of Chinese bureaucratic advancement, with multiple concurrent and sequential positions across party, government, and state-owned enterprise sectors. Decomposing these complex career histories into discrete positional observations yields over 4,000 structured biographical entries in the CPED benchmark. Each entry records a specific position with standardized fields: organization, location, role/title, start date, end date, and administrative rank.

5.2 Evaluation Design

To isolate coding performance, human and LLM coders are provided with an identical information environment. For each official, the input consists of the full Baidu Baike profile associated with that individual.[16] Under this setup, we generate two structured biographies per official. The baseline biography corresponds to the existing CPED record produced by trained RAs. The treatment biography is generated by applying a long-context LLM coder to the identical Baidu Baike text in a single pass. Because both biographies draw on the same evidence source and follow the same codebook, any performance differences can be attributed to the coder rather than to variation in information availability or task definition. We evaluate multiple LLM architectures (Grok-4.1-Fast, Gemini-2.5-Flash, and Qwen-2.5) to assess cross-model robustness.

CGT Construction While the CPED provides a high-quality human-coded benchmark, human annotation is not error-free. To establish a more reliable reference standard, we construct a *consolidated ground truth* (CGT) through a three-step validation pipeline. First, for each official, we pool all claims produced by both `Human_wiki` and `LLM_wiki` into a unified candidate set.[17] These

[16] In the Chinese context, official biographical information is highly standardized due to the party-state's institutionalized *nomenklatura* system. Baidu Baike, the dominant Chinese equivalent of Wikipedia, is the primary repository for official profiles. All sampled officials have Baidu Baike entries, which consolidate career narratives drawn from official announcements, government websites, and authoritative media sources. The human-coded ground truth in the CPED was originally derived primarily from these same Baidu Baike profiles.

[17] We use the subscript "wiki" as a generic shorthand to denote open-access, collaborative encyclopedia sources. For the Chinese sample, this refers specifically to Baidu Baike data; for the US and OECD samples discussed later, it refers to Wikipedia.

claims are then normalized into a standardized codebook (`entity, role, organization, start_date, end_date, status`), enabling direct comparison across coders. Second, each normalized claim is subjected to evidence-based validation using an LLM-as-judge protocol (Gu et al. 2024; H. Li et al. 2024), which evaluates supporting evidence from Baidu Baike and supplementary authoritative Chinese sources and classifies claims as *verified*, *contradicted*, or *uncertain*. Verified claims enter the CGT; contradicted claims are excluded; uncertain cases are flagged for review. Third, to assess the reliability of automated validation, we conducted a manual audit of 500 randomly sampled claims (50 officials × 10 claims). Two independent Chinese-speaking RAs reviewed the underlying evidence and judge classifications, achieving 94% agreement. Identified systematic error patterns were corrected by refining validation prompts and re-running affected cases. Full CGT construction procedures, judge prompts, and audit results are documented in Online Appendix A3.

Performance Metrics We evaluate coding performance at the official level by comparing system-generated claims against the CGT. Let $\widehat{C}_i$ denote claims produced by a candidate system for official i, and C_i^* denote CGT claims. We define true positives (TP_i), false positives (FP_i), and false negatives (FN_i) as follows:

$$\text{TP}_i = \left|\widehat{C}_i \cap C_i^*\right|, \quad \text{FP}_i = \left|\widehat{C}_i \setminus C_i^*\right|, \quad \text{FN}_i = \left|C_i^* \setminus \widehat{C}_i\right|.$$

From these quantities, we compute precision, recall, and F1 score:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \quad \text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}.$$

Precision captures the accuracy of extracted facts (the share of claims supported by verified evidence), while recall measures coverage (the share of true career events successfully recovered). **F1** is the harmonic mean balancing both dimensions. High precision but low recall yields incomplete biographies; high recall but low precision contaminates datasets with hallucinations.

Estimation Strategy To estimate differences in coding performance between human and LLM coders, we fit additive fixed-effect models of the following form:

$$Y_i = \alpha + \beta \cdot \mathbf{1}(\text{Coder}_i = \text{LLM}) + \gamma \cdot \text{Model}_i + \epsilon_i, \tag{1}$$

where $Y_i \in \{\text{F1}, \text{Precision}, \text{Recall}\}$ denotes the performance metric for official i. The indicator $\mathbf{1}(\text{Coder}_i = \text{LLM})$ captures whether the biography was produced by an LLM or by human coders, while Model_i includes fixed effects for LLM architecture (Grok, Gemini, Qwen). The coefficient β therefore identifies the average difference in coding performance between LLMs and humans, holding constant both the evidence source (Baidu Baike) and the extraction codebook (CPED). Standard errors are clustered at the official level, and 95% confidence intervals are obtained via nonparametric bootstrap (1,000 iterations).

5.3 Results

Figure 3 reports estimated differences in extraction performance relative to the human baseline, with coefficients from Equation 1 and 95% confidence intervals. Across all metrics, contemporary LLM coders match or exceed the human baseline when applied to the same curated Baidu Baike corpus. Grok-4.1-Fast increases F1 by 0.109 significantly, driven by improvements in both precision and recall. Gemini-2.5-Flash exhibits similar, though more moderate gains with balanced improvements in precision and recall. Even smaller open-source models such as Qwen-3 achieve near-human capability.

These results establish that coding accuracy is not a binding constraint in automated political-data production. When provided with curated inputs, LLMs can reliably map unstructured text to a complex, multi-field biographical codebook, achieving performance that matches or exceeds trained human coders. The most pronounced advantage lies in recall: leading models recover 10–16 percentage points more true career events than human coders working from the same corpus. This pattern is consistent with known limitations of manual annotation (principal–agent problems,

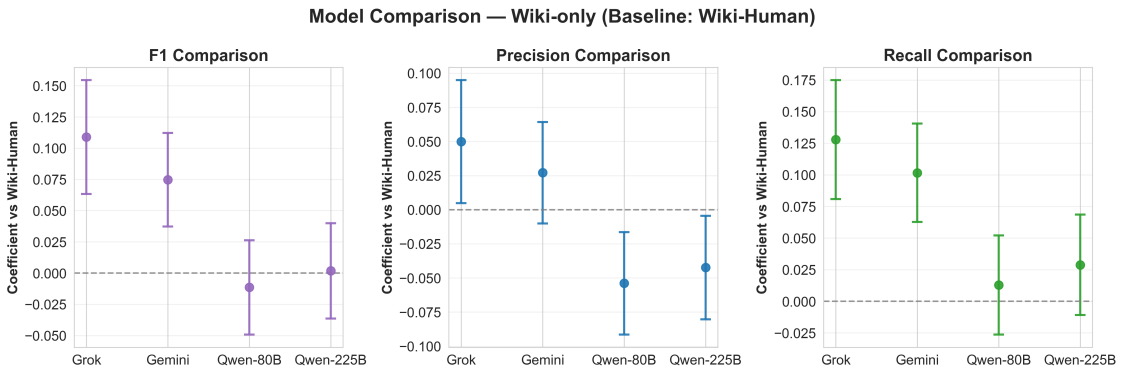

Figure 3: Experiment 1 Results: LLM coding performance relative to the human baseline (China sample, $N = 197$). Points indicate coefficient estimates with 95% confidence intervals. The human-coded baseline (`Human_wiki`) is normalized to zero. Positive values indicate that LLMs outperform human coders on the corresponding metric.

attention fatigue, selective reading, and time pressure), which lead human coders to systematically omit valid but less salient information, especially for officials with long and overlapping career histories. Importantly, these recall gains come with only modest changes in precision.

Greater differences exist in marginal production costs. For human coding, we assume a skilled coder paid $25 per hour. Given an average coding time of approximately 15 minutes per official, this yields a per-unit cost of $6.25. For LLM coding, costs are computed based on token-level pricing for long-context inference. A typical Baidu Baike biography contains approximately 5,000–9,000 input and output tokens combined. At the price of Gemini-2.5-flash, the best-performing model, each official costs on average $0.13 to process.

6 Experiment 2: The Synthesis Challenge

Having established that modern LLMs can validly and efficiently code curated biographical inputs, we turn to our central challenge: whether automated systems can match or exceed human collective curation (Wikipedia) in the upstream task of consolidating noisy and fragmented web sources into codeable evidence. We evaluate this challenge using two complementary settings: contemporary US political elites and ministerial officials from OECD countries. Unlike the Chinese case examined in Experiment 1, the US and OECD contexts lack comprehensive human-coded biographical

benchmarks covering the full range of elite career attributes.[18] Constructing such benchmarks manually would require thousands of RA hours and would reproduce precisely the scalability bottleneck that automated synthesis is designed to overcome. Rather than attempting to recreate human-coded benchmarks at prohibitive cost, Experiment 2 therefore evaluates synthesis performance in settings where only elite rosters are available ex ante and biographical information must be recovered from the open web. These contexts feature relatively high Wikipedia coverage, but with systematically varying degrees of information completeness and fragmentation beyond what encyclopedic curation captures. Together, they allow us to assess whether automated synthesis can recover biographical facts that are omitted, unevenly documented, or dispersed outside Wikipedia's curated summaries.

6.1 Data

US Political Elites The US sample comprises 198 contemporary political elites drawn from a comprehensive roster compiled by the authors.[19] We employ stratified random sampling focused on the post-2000 period to ensure high data density on the open web, selecting three equal cohorts of 66 officials: Cabinet members, state governors, and members of the 119th Congress. Unlike the centralized personnel records found in authoritarian hierarchies, American political data are structurally decentralized. Information is dispersed across federal databases, state archives, and local media, requiring the synthesis agent to navigate a highly heterogeneous source landscape to reconstruct coherent biographical narratives.

[18] Surprisingly, despite the size and maturity of the US political science literature, there is, to the best of our knowledge, no publicly available dataset providing CPED-style, career-long biographical coding for the full population of American political elites across offices and career stages. Existing resources typically focus on specific institutions (e.g., Congress or the presidency) or a narrow subset of attributes (Bonica 2016). For OECD countries, several cross-national databases document cabinet composition and tenure, but as shown in Table 1, the scope of coded attributes remains substantially narrower than the CPED, with limited coverage of education, pre-political careers, concurrent positions, and post-tenure trajectories.

[19] The full roster covers all state governors, members of Congress, Cabinet members, and Supreme Court justices from 1776 to 2025.

OECD Political Elites The OECD sample contains 200 ministerial officials from 36 member countries serving between 2011 and 2019, selected from the WhoGov 2.0 database (Nyrup and Bramwell 2020) via simple random sampling.[20] This sample tests the system's ability to handle linguistic and institutional breadth. The primary challenge is not merely depth, but the unevenness of digital curation: officials from smaller member states or minor portfolios often lack English-language Wikipedia entries, forcing the agent to retrieve and synthesize information from native-language government websites, party manifestos, and local press.

6.2 Evaluation Design

Experiment 2 isolates the contribution of upstream synthesis by holding the downstream LLM coder fixed and varying only the method of constructing evidentiary corpora. All biography types are ultimately coded by the same LLM using an identical extraction codebook; differences in performance therefore reflect variation in how evidence is located, consolidated, and curated prior to coding.

Synthesis Conditions We compare three synthesis conditions differing in how biographical evidence is assembled from the open web.

Human collective synthesis (Wikipedia baseline). Wikipedia represents the outcome of large-scale human collective curation: volunteer editors identify sources, resolve contradictions through citation norms, and compress verified information into narrative biographies. For this baseline condition, we use the existing Wikipedia page for each official [21] as the sole input corpus. This setting reflects a best-case benchmark for human synthesis, benefiting from years of accumulated editorial effort.

Agentic synthesis (full web). The full-web agent implements the iterative retrieval–reasoning loop described in Section 4. Starting from a broad query (official name and approximate role), the

[20] Unlike the China and US samples, which use stratified sampling to capture vertical hierarchies, the OECD sample uses horizontally comparable cabinet-level officials to maximize cross-national coverage.

[21] For wiki resources, we use all sources with "wiki" domain; we also exclude Grokipedia resources in all agent experiments.

agent issues successive searches conditioned on information discovered in earlier steps, inspects retrieved documents, and consolidates verified claims into a running synthetic report. Wikipedia is treated as one source among many rather than an authoritative endpoint. Each claim in the report is explicitly linked to archived sources, and the agent terminates once sufficient evidence has been gathered to populate the extraction codebook. On average, the agent conducts 15–25 searches and 12–20 document inspections per official, with synthesis costs of approximately $0.20 per case (search APIs), in addition to downstream coding costs (Search behavior and token usage across model families are summarized in Tables A2.1 and A2.2in the Online Appendix).

Agentic synthesis without Wikipedia. To assess whether agentic gains depend on access to curated encyclopedic content, we implement a non-Wikipedia variant in which all wiki-domain URLs are blocked during retrieval and document inspection. The agent must reconstruct a Wikipedia-equivalent evidentiary base entirely from non-Wikipedia sources, such as government websites, parliamentary records, party materials, and news archives. This condition directly tests whether automated synthesis can mitigate information-structure bias in contexts where encyclopedic curation is sparse or absent.

Biography Types Combining these synthesis conditions with a fixed downstream LLM coder yields three biography types. The baseline biography (`LLM_wiki`) is generated by applying the LLM coder to the Wikipedia page. The two treatment biographies are generated by applying the same coder to synthetic reports produced by the full-web agent (`LLM_agent`) and the non-Wikipedia agent (`LLM_nowiki`), respectively. Because the coder and extraction codebook are identical across conditions, performance differences isolate the contribution of upstream synthesis. Table 3 summarizes the experimental contrasts and costs.

Ground Truth Construction Evaluation in Experiment 2 relies on a similar design to Experiment 1, constructing CGT through evidence-based validation. For each official, we pool all claims extracted across synthesis conditions into a unified candidate set, normalize them into a common codebook, and validate each claim against archived source evidence using an automated judge.

Table 3: Experiment 2: Synthesis Method Comparison

Biography Type	Synthesis Method	Downstream Coder	Corpus Type	Length	Cost
`LLM_wiki` (baseline)	Human (Wikipedia)	LLM (fixed)	Wiki page	~8k	$0.01
`LLM_agent` (treatment 1)	Agent (full-web)	LLM (fixed)	Synthetic report	~10k	$0.21
`LLM_nowiki` (treatment 2)	Agent (non-wiki)	LLM (fixed)	Synthetic report	~10k	$0.21

Notes. All three biography types use the same downstream LLM coder (Grok-4.1-Fast). The experimental contrast isolates the synthesis contribution by holding the coder constant and varying only the upstream evidence construction method. Cost includes synthesis (search API) and coding (LLM API) expenses per official.

Verified claims constitute the CGT used for evaluation. To assess reliability, we conduct targeted manual audits on a random subset of claims across both the US and OECD samples, including multilingual cases. Agreement between automated judgments and human review exceeds 90%. Detailed consistency metrics for a randomly selected sample of officials across diverse linguistic contexts (OECD) are reported in Online Appendix Table A3.1.

Performance Metrics Performance is evaluated using the same individual-level precision, recall, and F1 metrics as in Experiment 1, computed by comparing system-generated claim sets against the CGT.

Estimation Strategy We estimate the effect of agentic synthesis using the following specification:

$$Y_i = \alpha + \beta_1 \cdot \mathbf{1}(\text{Synthesis}_i = \text{Agent}) + \beta_2 \cdot \mathbf{1}(\text{Synthesis}_i = \text{NoWiki}) + \gamma \cdot \text{Controls}_i + \epsilon_i, \quad (2)$$

where $Y_i \in \{\text{F1}, \text{Precision}, \text{Recall}\}$ denotes performance for official i. The baseline category is Wikipedia-based synthesis. Controls include downstream model indicators and sample fixed effects (US vs. OECD). Coefficients β_1 and β_2 capture the average performance difference between agentic and human collective synthesis, holding the coder constant. Standard errors are clustered at the official level, with 95% confidence intervals obtained via bootstrap.

6.3 Results

To rigorously quantify the synthesis challenge in these contexts, we first present the composition of URLs retrieved by our agentic framework across all three regions (including the China sample from Experiment 1 for comparison). Table 4 summarizes the distribution of these sources (see also Figure A4.4 in the Online Appendix for a visual breakdown). The data reveal distinct structural cross-region divergences. While the average volume of retrieved URLs is consistent across samples ($\approx$ 21–22 URLs per official), the composition of evidentiary sources differs fundamentally. The China baseline exhibits a high concentration of state-sanctioned information, with 78.4% of evidence derived from journalism (43.4%) or official government sources (35.0%). In contrast, the US sample demonstrates a highly fragmented distribution. Reliance on official government sources drops to 26.9%, while civil society sources, including non-wiki databases (10.4%) and NGOs (8.9%), comprise a substantial portion of the evidence base, compared with negligible levels in China. The OECD sample occupies an intermediate position, balancing journalism (25.8%) and government sources (22.1%) with a significant reliance on wiki-based references (17.4%). This contrast confirms that for democratic elites, valid extraction requires synthesizing evidence from a broad, diverse spectrum of non-official sources.

Table 4: Retrieved URL Category Composition by Sample (Top 6 Categories)

Region	Avg URLs	Govt	Wiki	Journalism	Databases	NGO	Social
China	21.2	7.44 (35.0%)	2.07 (9.8%)	9.22 (43.4%)	0.05 (0.2%)	0.41 (1.9%)	0.13 (0.6%)
US	22.4	6.02 (26.9%)	2.58 (11.5%)	3.10 (13.8%)	3.18 (14.2%)	1.99 (8.9%)	0.63 (2.8%)
OECD	22.1	4.87 (22.1%)	3.02 (13.7%)	5.70 (25.9%)	1.63 (7.4%)	1.71 (7.8%)	1.15 (5.2%)

Notes. Values represent the average number of unique URLs retrieved per official by the agentic framework, with the category's share of total regional volume in parentheses. Source categories are classified as follows: "Govt" includes official government sources; "Wiki" includes Wikipedia and wiki-derived encyclopedias; "Journalism" covers news media outlets; "Databases" refers to non-wiki structured reference databases (e.g., VoteSmart, Ballotpedia); "NGO" includes advocacy groups and NGOs; "Social" includes personal or professional social media platforms.

Figure 4 reports the aggregate effects of agentic synthesis relative to the Wikipedia baseline, pooling the US and OECD samples ($N = 398$).[22] Across both contexts, agentic synthesis substan-

[22]For completeness, Appendix A4.2 reports parallel results for the China sample, where the encyclopedic baseline is already highly curated.

tially improves coverage while maintaining acceptable accuracy. The full-web agent (`LLM_agent`) increases F1 by 14.7 percentage points, driven primarily by large recall gains of 31.4 points. Precision declines modestly by 5.2 points. Even when prohibited from accessing Wikipedia, the non-wiki agent (`LLM_nowiki`) achieves sizable improvements, increasing F1 by 11.7 points and recall by 24.3 points, with a comparable precision reduction of 4.1 points. Crucially, the dominance of recall gains underscores that agentic synthesis primarily bridges information gaps left by human curation, significantly expanding the scope of biographical records beyond encyclopedic baselines.

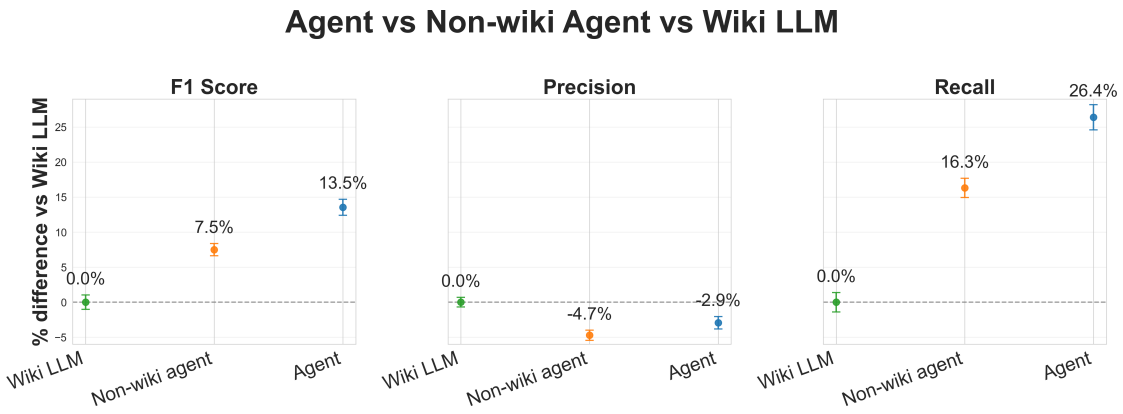

Figure 4: Agentic synthesis versus Wikipedia baseline (pooled US and OECD samples, $N = 398$). Points indicate coefficient estimates from Equation 2 with 95% confidence intervals. The Wikipedia baseline (LLM_wiki) is normalized to zero. Positive values indicate that agentic synthesis outperforms Wikipedia-based extraction.

The regional decomposition in Figure 5 further clarifies the magnitude and generality of these gains. Across regions, agentic synthesis raises absolute F1 from approximately 0.76–0.77 under the Wikipedia baseline to roughly 0.87–0.89, a performance level easily sufficient for downstream empirical applications. Importantly, these improvements are not driven by a single context. For the full-web agent, F1 increases by 15.1 percentage points in the US sample and by 14.3 points in the non-US sample, yielding a pooled gain of 14.7 points. Even when prohibiting Wikipedia access, the non-wiki agent delivers substantial improvements: F1 rises by 12.5 points in the US, 11.0 points in the non-US sample, and 11.7 points overall. Two implications follow. First, the scale of improvement is remarkably stable across regions, despite large differences in baseline coverage and information structure. Second, the slightly larger gains outside the US are consistent

with the intuition that agentic synthesis is most valuable where curated coverage is weakest. These results indicate that agentic workflows do not merely refine already well-documented cases but systematically elevate extraction quality across heterogeneous information environments.

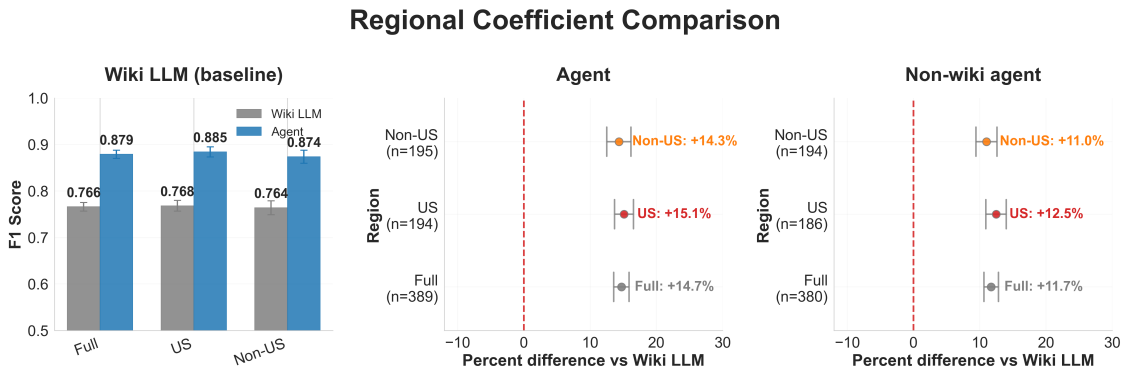

Figure 5: Agentic synthesis effects by sample. Points report F1 estimates with 95% confidence intervals. Baselines are normalized to sample-specific Wikipedia means (US: 0.82; OECD: 0.73).

Collectively, these results demonstrate that agentic synthesis systematically expands the informational scope of elite biographies. The significant gains in recall reflect a structural characteristic of human curation, which prioritizes high-visibility and prominent roles, even though existing Wikipedia data remain highly precise. Consequently, granular details such as early-career positions, officials from smaller nations, non-political affiliations, and concurrent appointments are frequently compressed or omitted. By actively querying local news archives, government records, and organizational filings, agentic workflows successfully recover this "long tail" of politically relevant information.

The resulting trade-off between precision and recall is both modest and favorable. While absolute precision remains robust at 0.82–0.85, recall improves substantially by 24–31 percentage points. For most empirical research, capturing a significantly larger volume of verified facts justifies a marginal increase in noise, especially given the transparency of source archives and the potential for downstream validation. Across both samples, agentic synthesis achieves F1 scores between 0.87 and 0.94. These levels exceed reported human intercoder reliability in comparable elite datasets and are achieved at a fraction of the cost of manual data collection.

7 Experiment 3: Why Does Synthesis Matter?

The first two experiments establish two results: LLMs can accurately code structured biographies when provided with curated evidence, and agentic retrieval can substantially expand coverage beyond Wikipedia. A natural follow-up question is whether explicit synthesis is still necessary once long-context models can ingest very large inputs. If a model can process hundreds of thousands of tokens, one might expect that simply concatenating all retrieved documents suffices for accurate extraction. This section evaluates that assumption. Holding the underlying web evidence fixed, we show that extraction performance depends materially on how evidence is represented to the coder. In particular, long context alone does not eliminate omission and degradation errors. Instead, a synthesis step that compresses and organizes evidence into a signal-dense representation is critical for reliable extraction.

7.1 Evaluation Design

To isolate representational effects, we hold the downstream coding procedure fixed and compare alternative representations of the same retrieved web evidence. Using a fixed Grok-based agent retrieval trajectory, we construct two long-context corpora from identical underlying sources (Table 5). The first is a raw internet corpus (`LLM_raw`), defined as the direct concatenation of all retrieved documents in full. The second is a refined internet corpus (`LLM_refined`), defined as a compressed, signal-dense representation comprising selected passages produced during the agentic reasoning loop. For comparison, we retain the Wikipedia-based long-context baseline (`LLM_wiki`). This design holds the evidence universe constant and varies only the representation presented to the coder.

We estimate biography-specific associations between recall and two mechanism proxies. The first captures a *quantity channel*: context length, operationalized using token-length bins. The second captures a *quality channel*: language composition, measured as the share of non-English

Table 5: Representations used in the diagnostic comparison

Condition	Upstream evidence construction	Downstream coder	Representation	Typical context length
`LLM_wiki`	Human (Wikipedia)	LLM	Wiki narrative	~8k
`LLM_raw`	Agent (fixed trajectory)	LLM	Raw concatenation	~300k
`LLM_refined`	Agent (fixed trajectory)	LLM	Refined passages	~30k

Notes. For `LLM_raw` and `LLM_refined`, the underlying retrieval trajectory is identical; only the representation supplied to the coder differs.

tokens in the coding input. Formally, we estimate specifications of the following form:

$$\text{Recall}_i = \alpha + \sum_{a \in \mathcal{A}} \mathbb{1}(\text{Bio}_i = a) \cdot (\psi_a M_i) + \gamma \cdot \text{Controls}_i + \epsilon_i,$$

where M_i denotes the mechanism proxy and ψ_a captures biography-specific slopes. This design allows us to diagnose whether long-context failures arise from scale effects, representation quality, or both.

7.2 Results

Figure 6 reports a clear refinement premium. Relative to the Wikipedia long-context baseline (`LLM_wiki`), the refined representation (`LLM_refined`) improves F1 by 10.4 percentage points and recall by 17.2 points, with only a modest precision change (-0.9 points). By contrast, raw concatenation (`LLM_raw`) yields substantially smaller gains: F1 increases by 4.5 points and recall by 8.8 points, accompanied by a larger precision decline (-2.8 points). Because `LLM_raw` and `LLM_refined` are constructed from the same retrieved evidence, this contrast isolates representation as the binding factor. Long-context extraction failures therefore stem not from missing evidence but from how evidence is organized and presented to the coder.

Figure 7 provides further mechanistic evidence. Panel B shows a monotonic quantity penalty: recall declines as context length increases beyond moderate ranges, with the largest losses in the longest bins, consistent with long-context omission errors. Panel A shows a complementary quality channel: higher non-English token shares are associated with lower recall in several bins, indicating that heterogeneous or weakly structured inputs further strain extraction. These patterns explain

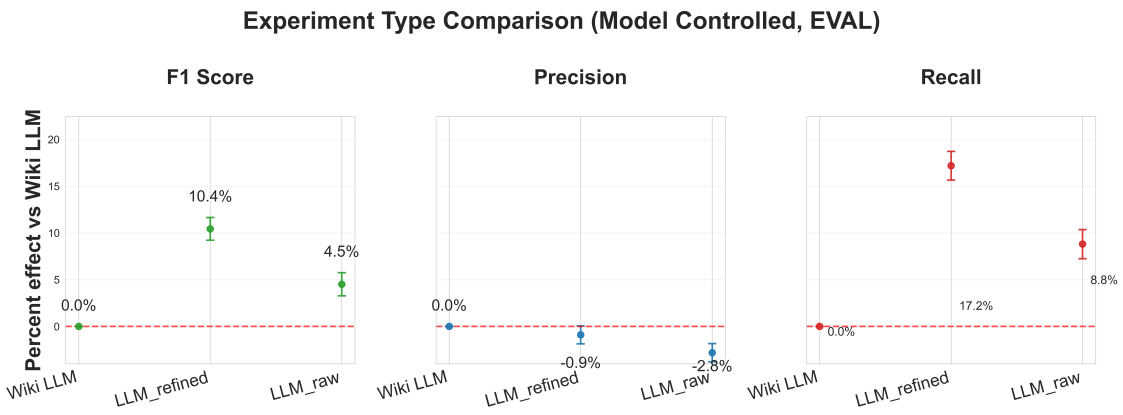

Figure 6: Refined versus raw corpora: the refinement premium.

why synthesis-then-coding outperforms raw concatenation even when both draw on the same evidence universe.

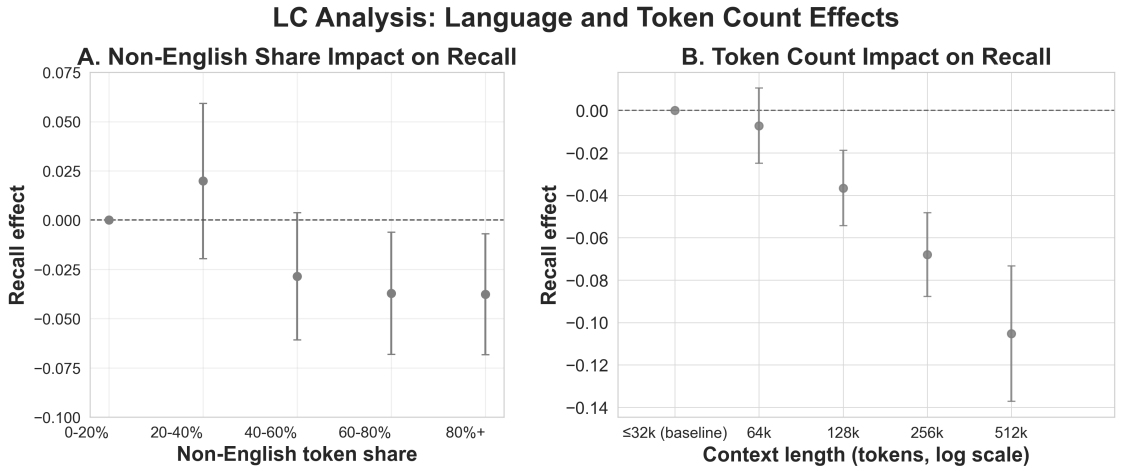

Figure 7: Mechanistic evidence on long-context extraction failures. Panel A plots associations between non-English token share and recall; Panel B plots associations between context length and recall. Points denote estimated contrasts relative to the baseline bin; bars indicate 95% confidence intervals.

These results clarify why explicit synthesis remains essential. Long context increases access to evidence, but does not guarantee effective use of that evidence. Without refinement, large inputs dilute signal, exacerbate attention limits, and amplify representational noise. Agentic synthesis mitigates these failures by compressing evidence into structured, signal-dense representations that align with the extraction task. In short, synthesis and long context are not substitutes for one another. Reliable large-scale extraction requires both. Model heterogeneity under long-context

conditions and descriptive comparisons of corpus composition are reported in Appendix A4.

8 Conclusion

"The historian," Carr (1961) famously observed, "is necessarily selective. The belief in a core of historical facts existing objectively and independently of the interpretation of the historian is a preposterous fallacy." A less noticed corollary applies with equal force to the political scientist: the structured datasets that undergird comparative inference are not neutral recordings of political reality but artifacts of particular production processes—constrained by labor costs, source availability, and the cognitive limits of coders. This paper takes that constraint seriously and asks whether modern language technology can relax it without sacrificing validity.

Our answer, based on three experiments spanning Chinese, American, and OECD political elites, is cautiously affirmative—with important qualifications about how automated systems must be designed to earn that optimism. When given curated biographical inputs identical to those used by trained RAs, contemporary LLMs match or exceed human coding quality, with leading models recovering 10–16 percentage points more verified career events than their human counterparts working from the same source. In open-web environments, where no curated Wikipedia biography exists, an agentic synthesis workflow raises absolute F1 scores from the mid-seventies to the high eighties—performance levels that meet or exceed reported human intercoder reliability in comparable elite datasets—at roughly three percent of the per-unit cost of manual collection. Holding the evidence universe fixed and vary only how that evidence is represented to the coder, we show that long-context concatenation is not a substitute for synthesis: raw document aggregation yields substantially smaller and noisier gains than an explicitly refined, signal-dense representation derived from the same retrieved sources.

These results carry complicated normative implications for a field that has long treated human coding as the unquestioned benchmark for complex extraction. First, the demonstration that LLMs can match and often exceed trained coders on identical inputs should prompt reflection

about whether the costs of manual annotation have been buying the validity they were assumed to guarantee. Human coding is subject to well-documented failure modes—principal–agent attrition, selective reading, attention fatigue, and inconsistent adjudication across coders—that systematic LLM evaluation does not share to the same degree. Second, however, automated pipelines introduce their own form of bias, which is less visible precisely because it is technical rather than human. Our mechanistic evidence shows that extraction performance degrades predictably under long inputs, multilingual corpora, and poorly structured source collections—biases that researchers may not notice if they do not inspect intermediate representations. This implies that validity depends on upstream design choices (e.g., retrieval strategy, evidence compression, source credibility weighting) that are no less consequential than the coder's codebook.

The broader significance of our framework lies in what it makes newly possible rather than merely what it does more cheaply. As Table 1 documents, the most analytically important datasets in comparative elite research have remained static for years or decades, updated only when sustained institutional funding and coordinated RA labor can be mobilized. This structural rigidity shapes inquiry: researchers generally study actors at the apex of formal institutions, in countries with dense English-language documentation, using attributes requiring the fewest contextual inferences. The result is a systematic information-structure bias: the distribution of available data drives theoretical attention rather than the reverse (Wilson and Knutsen 2022). An agentic synthesis workflow does not eliminate this bias, but it substantially attenuates two of its main sources—the cost of expanding coverage to lower-visibility elites and the difficulty of maintaining data currency as political environments evolve. That the non-Wikipedia agent delivers F1 gains nearly as large as the full-web agent suggests that the technology is most valuable precisely where human curation is thinnest: officials from smaller states, minor portfolios, or less-documented institutional contexts where the conventional approach would simply leave the field blank.

An interesting methodological parallel is worth noting. The precision–recall trade-off we document in agentic synthesis mirrors a well-known dilemma in human coding between breadth and accuracy. Time-pressured human coders tend to record the most salient positions and omit

earlier or concurrent roles that require more inferential effort to recover. The agentic workflow reverses this asymmetry: it excels at recovering the "long tail" of biographical facts precisely because its search behavior is anchored not by salience but by evidential completeness. The modest precision declines we observe are consistent with false positives generated at the margin, but they are interpretable, source-traceable, and addressable through downstream validation in a way that human omission errors typically are not.

Several directions follow naturally from these findings. The relationship between extraction and classification in political science remains underexplored. Many constructs typically treated as classification targets—regime type, populism, democratic backsliding, policy diffusion—could in principle be reconceptualized as aggregations of extracted sub-claims: time-stamped events, actor attributions, institutional changes. Testing whether a synthesis-then-coding architecture improves both the accuracy and auditability of such labels relative to direct prediction represents a promising frontier. A second direction concerns the contested end of the extraction spectrum. Our evaluations focus on relatively verifiable biographical facts where ground truth is unambiguous. Extending the framework to more interpretive attributes—rhetorical frames, policy positions, soft-power signals—raises harder questions about what "ground truth" means and how human and machine judgment should be combined when the target concept is itself contested. Finally, the biographical data produced by our framework create new empirical leverage for network approaches to elite politics. By standardizing career events, organizational affiliations, and temporal sequences across tens of thousands of officials, the resulting database permits reconstruction of fine-grained elite networks—co-service ties, overlapping tenures, shared educational institutions—at a scale and cross-national scope that hand-coded datasets could never sustain. How these networks shape recruitment, policy coalitions, and regime stability remains largely unexplored in comparative work, not because the theoretical questions are unimportant but because the data have not existed to answer them. They now can.

References

Alexiadou, D. (2015). "Ideologues, partisans, and loyalists: Cabinet ministers and social welfare reform in parliamentary democracies." In: *Comparative Political Studies* 48.8, pp. 1051–1086.

Alexiadou, D. (2022). "Cabinet ministers and inequality." In: *European Journal of Political Research* 61.2, pp. 326–350.

Alexiadou, D., W. Spaniel, and H. Gunaydin (2022). "When technocratic appointments signal credibility." In: *Comparative Political Studies* 55.3, pp. 386–419.

Angeli, G., M. J. J. Premkumar, and C. D. Manning (2015). "Leveraging linguistic structure for open domain information extraction." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 344–354.

Armstrong, B., T. D. Barnes, D. Chiba, and D. Z. O'brien (2024). "Financial crises and the selection and survival of women finance ministers." In: *American Political Science Review* 118.3, pp. 1305–1323.

Bäck, H., J. Teorell, A. Von Hagen-Jamar, and A. Quiroz Flores (2021). "War, performance, and the survival of foreign ministers." In: *Foreign Policy Analysis* 17.2, oraa024.

Bail, C. A. (2024). "Can Generative AI improve social science?" In: *Proceedings of the National Academy of Sciences* 121.21, e2314021121.

Baturo, A. (2016). "Cursus Honorum: Personal background, careers and experience of political leaders in democracy and dictatorship—New data and analyses." In: *Politics and Governance* 4.2, pp. 138–157.

Baturo, A., N. Dasandi, and S. J. Mikhaylov (2017). "Understanding state preferences with text as data: Introducing the UN General Debate corpus." In: *Research & Politics* 4.2, p. 2053168017712821.

Baturo, A. and J. A. Elkink (2022). "What countries select more experienced leaders? The PolEx measure of political experience." In: *British Journal of Political Science* 52.3, pp. 1455–1464.

Baturo, A. and J. Tolstrup (2023). "Incumbent takeovers." In: *Journal of Peace Research* 60.2, pp. 373–386.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). "Crowd-sourced text analysis: Reproducible and agile production of political data." In: *American Political Science Review* 110.2, pp. 278–295.

Benoit, K., S. De Marchi, C. Laver, M. Laver, and J. Ma (2025). "Using large language models to analyze political texts through natural language understanding." In: *American Journal of Political Science*.

Binderkrantz, A. S., J. G. Christensen, P. M. Christiansen, M. K. Nielsen, and H. H. Pedersen (2024). "Closed shutters or revolving doors? Elite career track similarity and elite sector transfers in Denmark." In: *European Journal of Political Research* 63.3, pp. 1022–1041.

Bomprezzi, P., A. Dreher, A. Fuchs, T. Hailer, A. Kammerlander, L. C. Kaplan, S. Marchesi, T. Masi, C. Robert, and K. Unfried (2025). *Wedded to Prosperity? Informal Influence and Regional Favoritism*. CEPR Discussion Paper 18878 (v.2).

Bonica, A. (2016). "Database on ideology, money in politics, and elections: Public version 2.0 [computer file]." In: *URL: https://data. stanford. edu/dime*.

Braun, M. and C. Raddatz (2010). "Banking on politics: When former high-ranking politicians become bank directors." In: *The World Bank Economic Review* 24.2, pp. 234–279.

Carnes, N. (2024). *White-collar government: The hidden role of class in economic policy making*. University of Chicago Press.

Carozzi, F. and L. Repetto (2016). "Sending the pork home: Birth town bias in transfers to Italian municipalities." In: *Journal of public economics* 134, pp. 42–52.

Carr, E. H. (1961). *What is History?* Cambridge, UK: Cambridge University Press.

De Luca, G., R. Hodler, P. A. Raschky, and M. Valsecchi (2018). "Ethnic favoritism: An axiom of politics?" In: *Journal of Development Economics* 132, pp. 115–129.

Dreher, A., M. J. Lamla, S. M. Lein, and F. Somogyi (2009). "The impact of political leaders' profession and education on reforms." In: *Journal of Comparative Economics* 37.1, pp. 169–193.

Ellis, C. M., M. C. Horowitz, and A. C. Stam (2015). "Introducing the LEAD data set." In: *International Interactions* 41.4, pp. 718–741.

Eschenauer-Engler, T. and B. Herre (2023). "Coup leaders: A new comprehensive dataset, 1950–2020." In: *European Political Science*. Forthcoming.

Fang, H., M. Li, and G. Lu (2025). *Decoding China's Industrial Policies*. Tech. rep. National Bureau of Economic Research.

Fearon, J. D., K. Kasara, and D. D. Laitin (2007). "Ethnic minority rule and civil war onset." In: *American Political Science Review* 101.1, pp. 187–193.

Fisman, R., J. Shi, Y. Wang, and W. Wu (2020). "Social ties and the selection of China's political elite." In: *American Economic Review* 110.6, pp. 1752–1781.

Fuchs, A. and K. Richert (2018). "Development minister characteristics and aid giving." In: *European Journal of Political Economy* 53, pp. 186–204.

Funke, M., M. Schularick, and C. Trebesch (2023). "Populist leaders and the economy." In: *American Economic Review* 113.12, pp. 3249–3288.

Gentzkow, M., B. Kelly, and M. Taddy (2019). "Text as data." In: *Journal of Economic Literature* 57.3, pp. 535–574.

Gerring, J., E. Oncel, K. Morrison, and D. Pemstein (2019). "Who rules the world? A portrait of the global leadership class." In: *Perspectives on politics* 17.4, pp. 1079–1097.

Gilardi, F., M. Alizadeh, and M. Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks." In: *Proceedings of the National Academy of Sciences* 120.30, e2305016120.

Goemans, H. E., K. S. Gleditsch, and G. Chiozza (2009). "Introducing Archigos: A dataset of political leaders." In: *Journal of Peace research* 46.2, pp. 269–283.

Grimmer, J. and B. M. Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." In: *Political analysis* 21.3, pp. 267–297.

Gu, J., X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al. (2024). "A survey on llm-as-a-judge." In: *The Innovation*.

Hallerberg, M. and J. Wehner (2012). "The educational competence of economic policymakers in the EU." In: *Global Policy* 3, pp. 9–15.

Halterman, A. and K. A. Keith (2024). "Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts." In: *arXiv preprint arXiv:2407.10747*.

Hassan, T. A., S. Hollander, L. Van Lent, and A. Tahoun (2019). "Firm-level political risk: Measurement and effects." In: *The quarterly journal of economics* 134.4, pp. 2135–2202.

Herre, B. (2023). "Identifying ideologues: A global dataset on political leaders, 1945–2020." In: *British Journal of Political Science* 53.2, pp. 740–748.

Jiang, J. (2018). "Making bureaucracy work: Patronage networks, performance incentives, and economic development in China." In: *American Journal of Political Science* 62.4, pp. 982–999.

Jiang, J. and M. Zhang (2020). "Friends with benefits: Patronage networks and distributive politics in China." In: *Journal of Public Economics* 184, p. 104143.

King, G., J. Pan, and M. E. Roberts (2013). "How censorship in China allows government criticism but silences collective expression." In: *American political science Review* 107.2, pp. 326–343.

Landry, P. F. (2008). *Decentralized Authoritarianism in China: the Communist Party's control of local elites in the post-Mao era*. Cambridge University Press.

Lee, D. S. and C. T. McClean (2022). "Breaking the cabinet's glass ceiling: the gendered effect of political experience in presidential democracies." In: *Comparative Political Studies* 55.6, pp. 992–1020.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In: *NeurIPS*.

Li, H., Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu (2024). "Llms-as-judges: a comprehensive survey on llm-based evaluation methods." In: *arXiv preprint arXiv:2412.05579*.

Licht, A. A. (2022). "Introducing Regular Turnover Details, 1960–2015: A dataset on world leaders' legal removal from office." In: *Journal of Peace Research* 59.2, pp. 277–285.

Liebman, B. L., M. E. Roberts, R. E. Stern, and A. Z. Wang (2020). "Mass digitization of Chinese court decisions: How to use text as data in the field of Chinese law." In: *Journal of Law and Courts* 8.2, pp. 177–201.

Lin, R. (2020). "The rise of technocratic leadership in the 1990s in the People's Republic of China." In: *Politics and Governance* 8.4, pp. 157–167.

Liu, H. (2024). "Meritocracy as authoritarian co-optation: Political selection and upward mobility in China." In: *American Political Science Review* 118.4, pp. 1856–1872.

Liu, N. F., K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang (2024). "Lost in the middle: How language models use long contexts." In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173.

Mallen, A., A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi (2023). "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822.

Mattes, M., B. A. Leeds, and N. Matsumura (2016). "Measuring change in source of leader support: The CHISOLS dataset." In: *Journal of Peace Research* 53.2, pp. 259–267.

Muthiah, S., B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan (2015). "Planned protest modeling in news and social media." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 2, pp. 3920–3927.

Nyrup, J. and S. Bramwell (2020). "Who governs? A new global dataset on members of cabinets." In: *American Political Science Review* 114.4, pp. 1366–1374.

Nyrup, J., C. H. Knutsen, P. E. Langsæther, and I. L. Kristiansen (2025). "Paths to power: A new dataset on the social profile of governments." In: *British Journal of Political Science* 55, e129.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott (2025). "How to train your stochastic parrot: Large language models for political texts." In: *Political Science Research and Methods* 13.2, pp. 264–281.

Palmer, A., N. A. Smith, and A. Spirling (2024). "Using proprietary language models in academic research requires explicit justification." In: *Nature Computational Science* 4.1, pp. 2–3.

Putnam, R. D. (1976). *The Comparative Study of Political Elites*. Englewood Cliffs, NJ: Prentice-Hall.

Raleigh, C. and D. Wigmore-Shepherd (2022). "Elite coalitions and power balance across African regimes: introducing the African cabinet and political elite data project (ACPED)." In: *Ethnopolitics* 21.1, pp. 22–47.

Reuter, O. J. and D. Szakonyi (2019). "Elite Defection under Autocracy: Evidence from Russia." In: *American Political Science Review* 113.2, pp. 552–568.

Ricart-Huguet, J. (2021). "Colonial education, political elites, and regional political inequality in Africa." In: *Comparative Political Studies* 54.14, pp. 2546–2580.

Shi, F., X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou (2023). "Large language models can be easily distracted by irrelevant context." In: *International Conference on Machine Learning*. PMLR, pp. 31210–31227.

Shih, V., C. Adolph, and M. Liu (2012). "Getting ahead in the communist party: explaining the advancement of central committee members in China." In: *American political science review* 106.1, pp. 166–187.

Svolik, M. W. (2012). *The politics of authoritarian rule*. Cambridge University Press.

Thrall, C. (2025). "Informational lobbying and commercial diplomacy." In: *American Journal of Political Science* 69.3, pp. 1147–1162.

Vittori, D., J.-B. Pilet, S. Rojon, and E. Paulis (2023). "Technocratic ministers in office in European countries (2000–2020): What's new?" In: *Political Studies Review* 21.4, pp. 867–886.

Wilson, M. C. and C. H. Knutsen (2022). "Geographical coverage in political science research." In: *Perspectives on Politics* 20.3, pp. 1024–1039.

Woldense, J. and A. Kroeger (2024). "Elite Change without Regime Change: Authoritarian Persistence in Africa and the End of the Cold War." In: *American Political Science Review* 118.1, pp. 178–194.

Wu, P. Y., J. Nagler, J. A. Tucker, and S. Messing (2023). "Large language models can be used to estimate the latent positions of politicians." In: *arXiv preprint arXiv:2303.12057*.

Yang, Z., P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning (2018). "HotpotQA: A dataset for diverse, explainable multi-hop question answering." In: *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao (2023). "ReAct: Synergizing reasoning and acting in language models." In: *International Conference on Learning Representations*.

Yu, S. and R. Jong-A-Pin (2020). "Rich or alive? Political (in)stability, political leader selection and economic growth." In: *Journal of Comparative Economics* 48.3, pp. 561–577.

Zhang, H. and J. Pan (2019). "Casm: A deep-learning approach for identifying collective action events with text and image data from social media." In: *Sociological Methodology* 49.1, pp. 1–57.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang (2024). "Can large language models transform computational social science?" In: *Computational Linguistics* 50.1, pp. 237–291.

Online Appendix

Table of Contents

A1 Formalizing Classification versus Extraction

To formalize the distinction emphasized in the main text, we represent (i) classification as a closed-set mapping and (ii) political fact extraction as a two-stage pipeline that couples evidence synthesis with structured record construction. The key implication is methodological: larger or "more capable" language models need not yield reliable gains for extraction if the binding constraint is open-domain discovery, long-context integration, and multi-stage task execution under explicit budgets.

A1.1 Definitions and notation

We use *extraction* to denote the end-to-end task of producing structured political facts from unstructured corpora (a single document, a fixed document collection, or an open-ended document universe), following a predefined codebook. Within extraction, we distinguish *synthesis* as evidence acquisition and refinement (search/browse/source selection, filtering, and condensation) from *coding* as mapping a fixed, refined corpus into codebook-conformant structured records.

For classification, let x denote the input text (a document or fixed set of documents) from an input space $\mathcal{X}$, and let $\mathcal{Y}$ denote a pre-defined, finite set of disjoint labels (e.g., $\mathcal{Y} = \{\text{Left}, \text{Right}\}$). The classification task is a mapping

$$f_{\text{classification}} : \mathcal{X} \to \mathcal{Y},$$

often implemented by selecting the most likely label,

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \Pr(y \mid x).$$

Classification is therefore discriminative: the output space is fixed and known ex ante, and errors are primarily *mislabeling*.

A1.2 Extraction as synthesis → coding

Synthesis corresponds to evidence retrieval and refinement over a large, heterogeneous source universe (e.g., the open web). Let $\mathcal{D}$ denote the (implicit) universe of candidate documents and let q denote a query (e.g., an entity name plus accumulated context from prior steps). Evidence retrieval can be written as a retrieval mapping that selects a bounded subset of evidence:

$$f_{\text{retrieval}} : (q, h) \to \mathcal{D}_k \subseteq \mathcal{D},$$

where h is interaction history and k is a number of retrieved corpus (or retrieved corpus). In practice, synthesis also includes filtering and condensation of $\mathcal{D}_k$ into a curated corpus that is feasible for downstream coding. Let g denote a condensation operator that maps retrieved evidence into a curated corpus x (e.g., a synthetic report) in an input space $\mathcal{X}$:

$$g : \mathcal{D}_k \to \mathcal{X}.$$

We therefore write synthesis as the composition

$$f_{\text{syn}}(q, h) = g(f_{\text{ret}}(q, h)) \in \mathcal{X}.$$

Coding then maps a fixed, refined corpus into structured records under an explicit codebook. Here, syn means synthesis. Let $\mathcal{B}$ denote a target codebook consisting of fields $\{b_1, \ldots, b_K\}$ (e.g., Organization, Role, Start Date, End Date). Let $\mathcal{V}$ denote the vocabulary of the language model and let $\mathcal{V}^{\leq L}$ denote the set of token sequences up to length L (a convenient representation for LLM outputs). A codebook-conformant record is a tuple $z = (v_1, \ldots, v_K)$ where each field value $v_k \in \mathcal{V}^{\leq L}$. Let $\mathcal{Z} \subseteq (\mathcal{V}^{\leq L})^K$ denote the set of such records, and let $\mathcal{T}$ denote the set of finite sequences of records (trajectories). We write the coding step as

$$f_{\text{code}} : \mathcal{X} \times \mathcal{B} \to \mathcal{T}.$$

The key difference from classification is that the output space is effectively open and the task is not separable: values are not drawn from a small closed set, and field-level decisions depend on other fields and on evidence scattered across documents. For elite biographies, the target is an ordered career trajectory rather than a single record. Let the extracted trajectory be $\hat{\tau} = (z_1, \ldots, z_T) \in \mathcal{T}$, where each $z_t \in \mathcal{Z}$ is a codebook-conformant record. End-to-end extraction is the two-stage composition:

$$\hat{\tau} = f_{\text{ext}}(q, h, \mathcal{B}) := f_{\text{code}}\Big(f_{\text{syn}}(q, h), \mathcal{B}\Big).$$

Accordingly, beyond mislabeling, extraction failures include hallucination, span/value errors, missing events (recall loss), and codebook violations.

A1.3 Summary

Table A1.1: Conceptual distinction between classification and extraction tasks.

Dimension	Classification (f_{cls})	Extraction (Synthesis → Coding)
Pipeline	Single-step mapping	Two-stage: synthesis f_{syn} (retrieval f_{ret} + condensation g) then coding f_{code}; end-to-end f_{ext}
Output space	$\mathcal{Y}$ (finite, closed)	Intermediate evidence $\mathcal{D}_k \subseteq \mathcal{D}$, refined corpus $x \in \mathcal{X}$, and trajectory $\hat{\tau} \in \mathcal{T}$ (open, effectively unbounded)
Objective	Label selection	Evidence synthesis + reconstruction (recover structured values)
Typical errors	Mislabeling	Missed facts (recall loss); Unsupported facts (precision loss)

A2 Architecture and Model Details

This appendix documents the system architecture and model configuration used to produce the agentic biographies and the derived long-context corpora used in the experiments. In the main text, we emphasize the core conceptual feature—a tool-using, ReAct-style workflow for iterative evidence gathering and synthesis (Yao et al. 2023). Here we provide additional implementation detail to make the design auditable and to clarify what is held constant across comparisons. The system architecture is illustrated in Figure 2 in the main text.

A2.1 Architecture and information flow

We implement a Supervisor–Worker architecture to manage the cognitive overhead of open-web synthesis. The central division of labor is between (i) strategic, long-horizon reasoning about what is missing and what to search next and (ii) tactical, short-horizon retrieval and reading of specific sources.

Core components. The **Supervisor** interprets the extraction objective and codebook, orchestrates the workflow over multiple cycles, and produces the final consolidated record. The Supervisor does not process the full raw web corpus directly; instead, it operates on structured evidence packets returned by specialized **Searcher** workers. The **Archive** stores retrieved content and provenance metadata (e.g., URLs and retrieval time), enabling deduplication, backtracking when contradictions arise, and transparent linkage between claims and supporting passages. The **Coder** converts the Supervisor's stabilized draft and the archived evidence bundle into the structured biography output used for evaluation.

Two-stage pipeline. Our pipeline separates the extraction of political facts into (i) **synthesis**, which transforms a large, noisy document set into a higher-signal evidentiary representation, and (ii) **coding**, which maps that representation into a structured biography under a fixed codebook. This separation is crucial for interpretation: it allows us to hold the downstream coder constant while varying the upstream synthesizer (RQ2), and to isolate representation effects while holding the underlying web evidence fixed (Section 7).

Intermediate data structures. To keep multi-cycle synthesis auditable, we use explicit intermediate objects: (i) a **Structured Input** (objective, codebook, and constraints), (ii) a **System State** (running plan, partial biography, unresolved gaps, and bookkeeping), (iii) **Information Batch Overviews** produced by Searchers (source details, task-specific summaries, and extracted passages), and (iv) a **Structured Final Output** (codebook-conformant biography with evidence pointers via the Archive). Prompts are role-specific and enforce these interfaces; full prompt templates and tool specifications are provided in the replication materials.

A2.2 Models and operational constraints

We evaluate multiple model families as downstream coders (e.g., `LLM_wiki`) and as agentic components (e.g., `LLM_agent`). The primary model families used in this version are **Grok-4.1-Fast**,

Gemini-2.5-Flash, **Qwen-3-80B**, and **Qwen-3-225B**. Across experiments, we hold prompts and codebooks fixed within each role and keep system budgets (e.g., maximum steps and termination rules) constant within comparison Bios.

Model choice is consequential in agentic synthesis because end-to-end performance depends not only on reading comprehension, but also on tool-use reliability and the ability to sustain long-horizon interaction without drifting. We therefore prioritize models that jointly satisfy three practical constraints: strong reasoning in a fixed corpus, robust multi-turn planning and tool use, and affordability at scale. These constraints motivate the inclusion of "Fast/Flash" variants where available and efficient open-weight alternatives.

Finally, we design synthesis to remain operationally bounded. The Supervisor maintains a running search summary and a gap list, decomposes the task into Searcher instructions, and terminates when a step budget is reached or when the marginal value of additional retrieval declines. Because the workflow logs its actions and preserves archived evidence, we can audit intermediate representations and support claim verification during CGT construction (Appendix A3).

Table A2.1: Agent Search Metrics by Model and Region

Model	Version	Region	Officials	Searched (Avg)	Search Times (Avg)	Used URLs (Avg)
Gemini 2.5 Flash						
Gemini 2.5 Flash	non_wiki	Overall	398	62,389 (156.75)	8,935 (22.44)	9,299 (23.37)
Gemini 2.5 Flash	non_wiki	US	198	30,672 (154.91)	4,220 (21.31)	4,917 (24.83)
Gemini 2.5 Flash	non_wiki	OECD	200	31,717 (158.59)	4,715 (23.57)	4,382 (21.91)
Gemini 2.5 Flash	model_wiki	Overall	398	61,046 (153.35)	8,035 (20.18)	7,075 (17.78)
Gemini 2.5 Flash	model_wiki	US	198	28,917 (146.05)	3,651 (18.44)	3,348 (16.91)
Gemini 2.5 Flash	model_wiki	OECD	200	32,129 (160.65)	4,384 (21.92)	3,727 (18.64)
Grok-4.1-Fast						
Grok-4.1-Fast	non_wiki	Overall	398	68,677 (172.60)	5,706 (14.34)	10,020 (25.18)
Grok-4.1-Fast	non_wiki	US	198	36,218 (182.92)	2,992 (15.11)	5,020 (25.35)
Grok-4.1-Fast	non_wiki	OECD	200	32,459 (162.29)	2,714 (13.57)	5,000 (25.00)
Grok-4.1-Fast	model_wiki	Overall	598	95,260 (159.28)	5,938 (9.94)	13,045 (21.82)
Grok-4.1-Fast	model_wiki	US	198	30,230 (152.68)	2,249 (11.36)	4,433 (22.39)
Grok-4.1-Fast	model_wiki	China	200	36,735 (183.68)	1,522 (7.61)	4,224 (21.12)
Grok-4.1-Fast	model_wiki	OECD	200	28,295 (141.47)	2,167 (10.84)	4,388 (21.94)
Qwen						
Qwen	model_wiki	Overall	398	41,411 (104.06)	5,747 (14.43)	8,640 (21.71)
Qwen	model_wiki	US	198	21,225 (107.20)	2,788 (14.08)	4,646 (23.46)
Qwen	model_wiki	OECD	200	20,186 (100.93)	2,959 (14.79)	3,994 (19.97)
Qwen 225B						
Qwen 225B	model_wiki	Overall	398	35,302 (88.67)	4,521 (11.36)	7,121 (17.89)
Qwen 225B	model_wiki	US	198	16,626 (83.97)	2,204 (11.13)	3,285 (16.59)
Qwen 225B	model_wiki	OECD	200	18,676 (93.38)	2,317 (11.59)	3,836 (19.18)

Notes. This table presents detailed search metrics for the agent search process across different models, versions, and regions. Each metric column reports both the total count and the per-official average in parentheses. "Searched (Avg)" shows total search results retrieved (average per official). "Search Times (Avg)" shows the number of search operations performed (average per official). "Used URLs (Avg)" counts the total number of useful URLs retrieved by agent from search results (average per official). Overall region represents aggregated totals across all sub-regions (e.g., US + OECD, or US + China + OECD).

Table A2.1 documents the search behavior of different agent models during the upstream retrieval phase. For each model family (Gemini 2.5 Flash, Grok, Qwen, Qwen 225B), we report metrics across multiple configurations (non_wiki and wiki variants) and geographic regions (Overall, US, OECD, China). The "Searched" column captures the total volume of search results processed, while "Search Times" indicates how many search queries each model issued. "Used URLs" counts the distinct web sources each model successfully retrieved, providing a proxy for retrieval breadth. These metrics reveal substantial variation in search strategies across models: some models (e.g., Gemini) issue more queries and retrieve more results, while others (e.g., Grok) converge more efficiently on relevant sources.

Notably, Grok demonstrates superior search efficiency across all models. As shown in Table A2.1, Grok achieves comparable or better coverage with substantially fewer search operations: it requires only 11.7 search times per official on average (across both versions), compared to 21.3 for Gemini 2.5 Flash and 14.4 for Qwen models. This efficiency translates directly into lower resource consumption and cost—Grok's per-official cost ($0.15–$0.18) is approximately 4–5× lower than Gemini 2.5 Flash ($0.65–$0.79) while maintaining competitive retrieval quality (88–102 unique URLs per official versus 123 for Gemini). Grok's ability to converge on relevant sources with fewer queries makes it particularly well-suited for large-scale agentic synthesis where cost and efficiency are critical constraints.

Larger models (Qwen 225B) tend to generate more output tokens than smaller variants (Qwen 80B), driving up costs despite similar input prices. The 225B variant v2 generates 5.2× more output tokens than the 80B variant (29.6M vs. 5.7M), contributing to its 2.9× higher total cost.

Table A2.2 reports the token usage and costs for each model configuration across different modes. Agent costs represent the full agentic synthesis pipeline (Searcher + Supervisor + Coder). Wiki LC (long-context) costs represent single-pass coding from Wikipedia pages. LC Raw and LC Synth are long-context variants using retrieved documents.

A2.2.1 Other infrastructure costs

We relied on external providers for web research and robust web content retrieval. In particular, we used Jina and Exa as retrieval services capable of extracting full page contents from URLs. For web search, we used serp.dev to obtain programmatic access to Google's search results.

Beyond model inference costs, production deployments should reserve budget for several API-related expenses. For search, about $400 covers roughly 400,000 queries (around $0.001 per request). Large-scale document fetching via Jina and exa services is on the order of $200. Overall, our full experimental effort—including development, testing, experimentation, and evaluation—amounted to approximately $5,000 in total spending across all LLM APIs, search, and retrieval services, as well as coders' hiring cost.

Table A2.3 breaks down the average number of URLs retrieved per official by type, model, and region. We categorize URLs into eight types using a politician-centric reliability hierarchy: official government sources (primary/authoritative), wiki pages, news/journalism media (tertiary/interpretive), non-wiki reference databases (secondary/structured), social media platforms, NGO/advocacy sources, commercial sources, and other sources. The distribution varies substantially across models and regions. Notably, wiki variants consistently retrieve more wiki pages (2.16–4.17 per official) compared to non-wiki variants (1.36–2.04), while non-wiki variants rely more heavily on non-wiki reference databases (e.g., Grok non-wiki: 2.74 per official in US vs Grok v7 wiki: 2.34). Grok v7

Table A2.2: Model Token Usage and Costs

Model	Mode	Input (M)	Output (M)	Input ($)	Output ($)	Total ($)	Per Official ($)
Agent Models							
Grok 4 Fast	Agent Wiki	271.7	13.1	$54.35	$6.54	$60.88	$0.153
Grok 4 Fast	Non-Wiki	323.4	15.0	$64.68	$7.49	$72.17	$0.181
Gemini 2.5 Flash	Agent Wiki	754.5	12.3	$226.36	$30.79	$257.15	$0.646
Gemini 2.5 Flash	Non-Wiki	938.2	12.7	$281.46	$31.69	$313.15	$0.787
Qwen3 225B	Agent Wiki	517.4	29.6	$93.12	$15.99	$109.11	$0.274
Qwen3 80B	Agent Wiki	341.4	5.7	$30.72	$6.25	$36.97	$0.093
Long-Context (LC) Modes							
Grok 4 Fast	Wiki	9.0	1.3	$1.79	$0.64	$2.43	$0.006
Grok 4 Fast	LC Raw	51.2	1.4	$10.23	$0.69	$10.92	$0.027
Grok 4 Fast	LC Synth	23.4	1.8	$4.68	$0.88	$5.56	$0.014
Gemini 2.5 Flash	Wiki	13.0	0.6	$3.90	$1.59	$5.49	$0.014
Gemini 2.5 Flash	LC Raw	45.3	0.8	$13.59	$1.98	$15.57	$0.039
Gemini 2.5 Flash	LC Synth	25.3	0.8	$7.60	$2.04	$9.64	$0.024
Qwen3 225B	Wiki	13.0	2.5	$2.34	$1.37	$3.71	$0.009
Qwen3 225B	LC Raw	29.5	2.0	$5.31	$1.06	$6.37	$0.016
Qwen3 225B	LC Synth	24.3	2.4	$4.37	$1.28	$5.65	$0.014
Qwen3 80B	Wiki	12.9	0.5	$1.16	$0.51	$1.67	$0.004
Qwen3 80B	LC Raw	29.5	3.3	$2.65	$3.61	$6.26	$0.016
Qwen3 80B	LC Synth	24.4	4.1	$2.19	$4.48	$6.67	$0.017

Notes. Sample size: N=398 officials for Agent modes; N varies for LC modes. Agent Wiki includes models with Wikipedia access; Non-Wiki blocks Wikipedia during synthesis. Wiki represents single-pass long-context coding from Wikipedia pages only. LC Raw uses retrieved documents; LC Synth uses supervisor-enhanced retrieved documents. Input/Output in millions (M) of tokens. Price based on openrouter model prices.

(wiki) shows particularly high government source usage in China (7.44 per official, representing 35% of all URLs) compared to US (26.9%) and OECD (22.1%). China data reflects only the final Grok v7 experiment due to processing log loss for earlier China search results.

Table A2.3: URL Types per Official by Model and Region

Model	Region	Total	Govt	News	Wiki	Reference	Platforms	NGO	Commercial	Other
Grok										
Grok non-wiki	US	25.61	9.02	3.80	0.18	4.04	0.71	2.28	0.78	4.81
Grok non-wiki	OECD	25.00	7.56	7.27	0.34	1.94	1.25	2.12	1.28	3.24
Grok wiki	US	22.39	6.02	3.10	2.58	3.18	0.63	1.99	0.53	4.35
Grok wiki	China	21.23	7.44	9.22	2.07	0.05	0.13	0.41	0.00	1.92
Grok wiki	OECD	22.05	4.87	5.70	3.02	1.63	1.15	1.71	1.23	2.73
Gemini										
Gemini non-wiki	US	25.09	7.04	3.37	0.02	4.95	1.21	2.77	1.02	4.72
Gemini non-wiki	OECD	22.02	4.62	5.43	0.03	2.78	2.20	2.11	1.86	2.99
Gemini wiki	US	16.99	3.84	1.79	2.58	2.86	0.66	1.40	0.74	3.12
Gemini wiki	OECD	18.92	3.43	3.79	3.03	1.91	1.67	1.50	1.48	2.11
Qwen 225B										
wiki	US	16.85	4.59	2.24	2.65	2.35	0.83	1.58	0.19	2.41
wiki	OECD	19.18	5.18	4.76	2.48	1.37	1.69	1.25	0.46	1.99
Qwen 80B										
wiki	US	23.46	6.71	3.85	2.68	2.74	1.51	1.76	0.49	3.73
wiki	OECD	19.97	4.59	2.81	2.77	2.06	2.98	1.39	0.76	2.60

Notes. This table shows the average number of URLs retrieved per official by type, model, and region. Grok v7 = model_wiki variant; Grok non-wiki = non_wiki variant; Gemini v2 = model_wiki variant; Gemini non-wiki = non_wiki variant; Qwen 225B v2 and Qwen 80B = model_wiki variants. China data reflects only the final Grok v7 experiment due to processing log loss for earlier China search results. Govt = official government sources; News = journalism and media; Wiki = Wikipedia and wiki-style pages; Reference = non-wiki reference databases (e.g., VoteSmart, Ballotpedia); Platforms = social media platforms; NGO = advocacy/NGO sources; Commercial = commercial/business sources; Other = uncategorized sources including entertainment, media, search engines, and miscellaneous.

A3 Consolidated Ground Truth (CGT) Construction

This appendix provides the full claim-level CGT protocol summarized in the main text. Our goal is to produce a defensible, auditable reference set of claims for scoring.

A3.1 Protocol (pooling, consensus, and verification)

Inputs (fixed pool per individual). For each individual i, we construct a fixed pool of **9 biographies**: four agent biographies (two agent model families × two variants), four `LLM_wiki` biographies (four coder models applied to the same Wiki corpus), and one human-written Wiki

biography (`Human_wiki`). We construct the CGT from this pool and score all candidate systems against it. Long-context baselines (`LLM_raw` and `LLM_refined`) are scored against the CGT but are not included in the CGT pool to avoid mechanically altering the consensus set.

Step 1 (claim extraction and normalization). We decompose each biography into a set of atomic, comparable claims (e.g., education events; offices held with dates; party membership). We then normalize claims to reduce superficial disagreement: we canonicalize entity names and common aliases when available; standardize role and organization strings (e.g., ministry/agency names); and harmonize date formats, resolving partial dates into comparable intervals when possible. After normalization, paraphrases that express the same event are treated as the same claim.

Step 2 (consensus filter for high-confidence claims). For each normalized claim, we compute its presence rate in the 9-biography pool:

$$presence(\text{claim}) = \frac{\#\{\text{bios containing claim}\}}{9}.$$

Claims with $presence \geq 5/9$ enter the CGT as **high-confidence** claims. Claims with $presence \leq 4/9$ are treated as **disputed/low-confidence** and proceed to evidence verification.

Step 3 (evidence-conditional verification for low-confidence claims). Low-confidence claims are evaluated against a pooled evidence bundle: the union of archived passages and sources collected across all agent runs and variants for the same individual. We add a low-confidence claim to the CGT only if it is supported by explicit evidence in this pooled archive. Pooling across agent runs mitigates dependence on any single retrieval trajectory and improves robustness to idiosyncratic search failures. In our implementation, we operationalize this step with an evidence-conditional verifier (GPT-5-mini), which receives the candidate claim and the pooled archive text and returns a supported/unsupported judgment. In practice, we used the soft label to let LLMs label the level of support by 1-5, and treat claims scored above 3 as supported claims.

Step 4 (CGT definition and scoring). Let C_i^{High} denote the set of high-confidence claims and $C_i^{\text{Validated}}$ the set of evidence-validated low-confidence claims. We define the claim-level CGT as:

$$C_i^* = C_i^{\text{High}} \cup C_i^{\text{Validated}}.$$

Each candidate system output is converted into a normalized claim set $\widehat{C_i}$ and scored against C_i^* using precision, recall, and F1 as defined in the main text.

A3.2 Audit checks

We validate the reliability of the automated CGT construction through two distinct audit studies. To assess whether evidence-conditional verification aligns with expert judgment and external search verification, we drew a random sample of 20 officials from the OECD dataset. Table A3.1 summarizes the consistency rates, measured as the percentage of exact matches between the automated CGT verdicts and the alternative verification methods.

Human–machine alignment. Two graduate research assistants independently verified extracted claims for the sampled officials. Auditors were provided with the claim text and access to open-web search but were blinded to the model's verdict. For non-English sources, auditors utilized translation tools alongside original source inspection. Agreement rates were calculated by comparing human judgments against the automated judge's outputs. As shown in Table A3.1, the average agreement rate is 91.3%, indicating strong alignment between the automated verifier and human judgment in adjudicating low-consensus discoveries.

External validation (Exa). To rule out model-specific artifacts in the retrieval process, we cross-validated claims using Exa deepsearch, a neural search engine optimized for semantic retrieval and fact checking (Exa 2025). We queried Exa to retrieve high-quality, independent evidence for each claim and compared its verification results with our pipeline's verdicts. The analysis reveals a negligible discrepancy rate (average agreement 98.7%), confirming that the synthesized ground truth is factually grounded and robust to retrieval method variations.

Table A3.1: Audit results: Consistency checks across verification methods for sampled OECD officials.

			Agreement Rate (%)	
Country	Official Name	Position (Abbreviated)	Human	Exa
CZE	Pavel Blazek	Min. of Justice	93.5	98.1
CZE	Jaromir Drabek	Min. of Labor & Social Affairs	89.0	98.8
JPN	Aiko Shimajiri	Min. in Charge of "Cool Japan" Strategy	91.4	99.6
JPN	Kenichiro Sasae	Ambassador to the US	95.0	97.7
SVK	Frantisek Ruzicka	Permanent Rep. to the UN (NY)	92.1	99.4
SVK	Pavol Pavlis	Min. of Economy	92.4	98.4
DNK	Rasmus Prehn	Min. for Development Cooperation	94.9	99.7
DNK	Kirsten Brosbol	Min. of Environment	90.0	97.9
KOR	Ju Chul-Ki	Senior Sec. for Foreign Affairs & Security	88.3	98.4
KOR	Lee Byung-Ho	Dir., National Intelligence Service	90.5	99.6
COL	Alfonso Gomez Mendez	Min. of Justice & Law	88.2	98.0
COL	Luis Felipe Henao Cardona	Min. of Housing & Territorial Dev.	91.6	97.7
FIN	Jan Vapaavuori	Min. of Economic Affairs	89.8	99.2
FIN	Jari Lindstrom	Min. of Justice & Employment	94.0	99.3
SVN	Bostjan Zeks	Min. w/o Portfolio (Slovenians Abroad)	92.0	98.9
SVN	Gorazd Zmavc	Min. w/o Portfolio (Slovenians Abroad)	91.2	98.8
IRL	Anne Colette Anderson	Permanent Rep. to the UN (NY)	90.7	99.7
IRL	Katherine Zappone	Min. for Children & Youth Affairs	88.4	97.8

Note: Agreement Rate indicates the percentage of claims where the auditor (Human or Exa) reached the same verification verdict (Supported/Unsupported) as the automated CGT pipeline.

A4 Supplementary Results

This section presents additional figures and analyses referenced in the main text to support key empirical claims. We focus on: (1) model performance without external resources, (2) comparison of agent-synthesized and Wiki-based corpora, (3) diagnostic checks of model heterogeneity under long-context constraints, (4) corpus composition and compression, (5) cross-national heterogeneity in retrieved corpora composition, and (6) granular mechanism plots illustrating recall dynamics and language effects.

A4.1 Model performance without external resources

Figure A4.1 illustrates that models exhibit poor performance on both precision and recall when operating without external resource access (e.g., without web search or retrieved documents). This three-panel comparison highlights the substantial performance gap between models with and without access to external information sources.

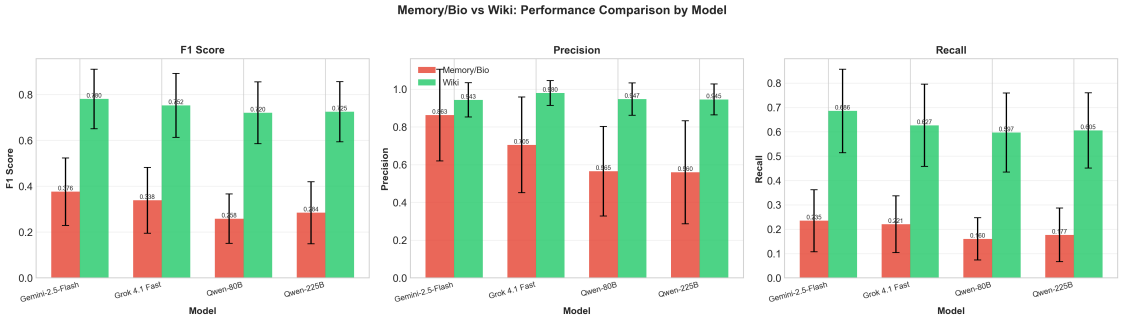

Figure A4.1: Three-panel comparison showing poor model performance without external resources on both precision and recall. Models with access to external resources (web search, retrieved documents) significantly outperform those operating without such access.

A4.2 Agentic Synthesis in a High-Curation Setting: China

As a supplementary analysis, we examine agentic synthesis performance in the China setting, where the encyclopedic baseline is unusually strong. Unlike the U.S. and OECD samples discussed in the main text, Chinese political elites benefit from highly standardized and centrally curated biographical documentation, and Baidu Baike entries are typically comprehensive. As a result, any gains from synthesis are expected to be modest if the agent primarily reproduces already well-documented information. Figure A4.2 reports the comparison between agent-synthesized biographies and the Wiki long-context baseline in this setting. Consistent with expectations, the magnitude of improvements is substantially smaller than in the U.S. and OECD samples. The agent increases F1 by 2.7 percentage points and recall by 3.8 points, accompanied by a modest precision gain of 1.4 points.

These results serve two purposes. First, they confirm that agentic synthesis does not degrade performance in environments where curated encyclopedic coverage is already strong. Second, the

presence of small but detectable recall gains indicates that even in highly curated contexts, synthesis can recover incremental information omitted from baseline entries, such as minor concurrent appointments or short transitional roles. Together, the China results reinforce the interpretation of the main findings: the value of agentic synthesis scales with gaps in existing curation, yielding large gains where coverage is incomplete and converging toward parity where high-quality encyclopedic resources already exist.

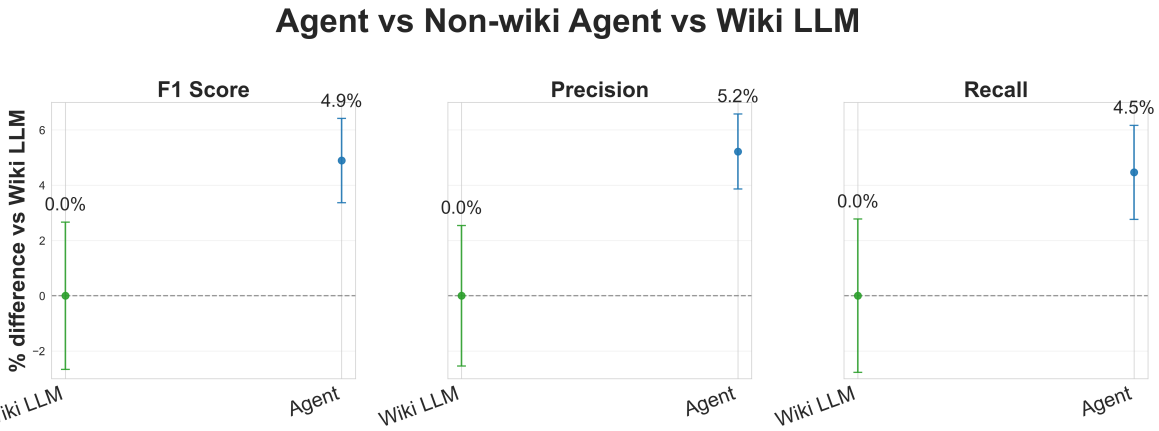

Figure A4.2: China setting: Comparison of agent-synthesized biographies and encyclopedic long-context baseline.

A4.3 Model heterogeneity under long-context conditions

We visualize performance disparities between coder models under uniform, long-context scenarios. Figure A4.3 documents how both raw and refined pipelines (`LLM_raw` and `LLM_refined`) are affected.

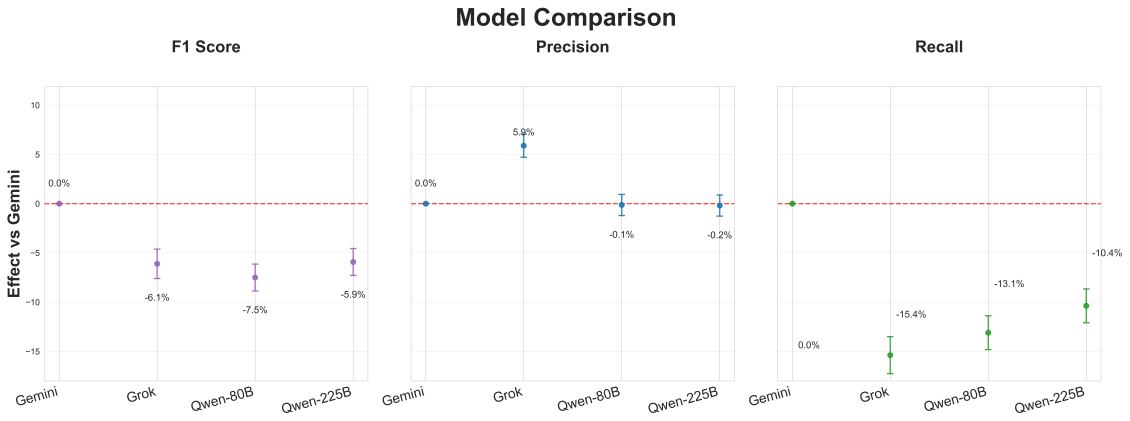

Figure A4.3: Performance differences across coder models (experiment-controlled) under long-context conditions, including both `LLM_raw` and `LLM_refined` variants.

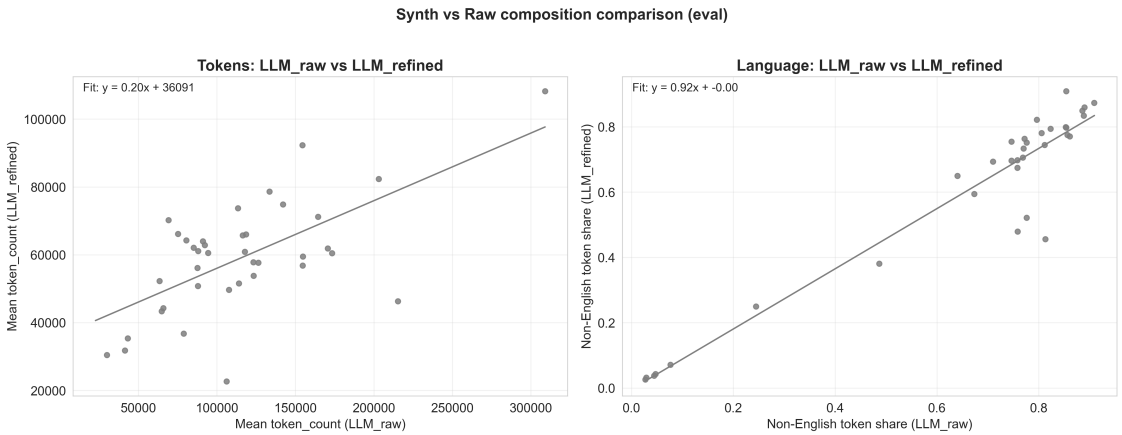

Figure A4.4: Token length and language composition for raw vs. refined corpora: refinement compresses token count while retaining high compositional correspondence.

A4.4 Cross-national heterogeneity in retrieved corpora composition

Figures A4.4, and A4.5 illustrate cross-national heterogeneity in retrieved corpora composition. Figure A4.4 shows that refinement compresses token count while retaining high compositional correspondence between raw and refined corpora. Figure A4.5 shows the percentage of non-English resources in each country's retrieved corpus, illustrating substantial variation in language diversity across the OECD sample.

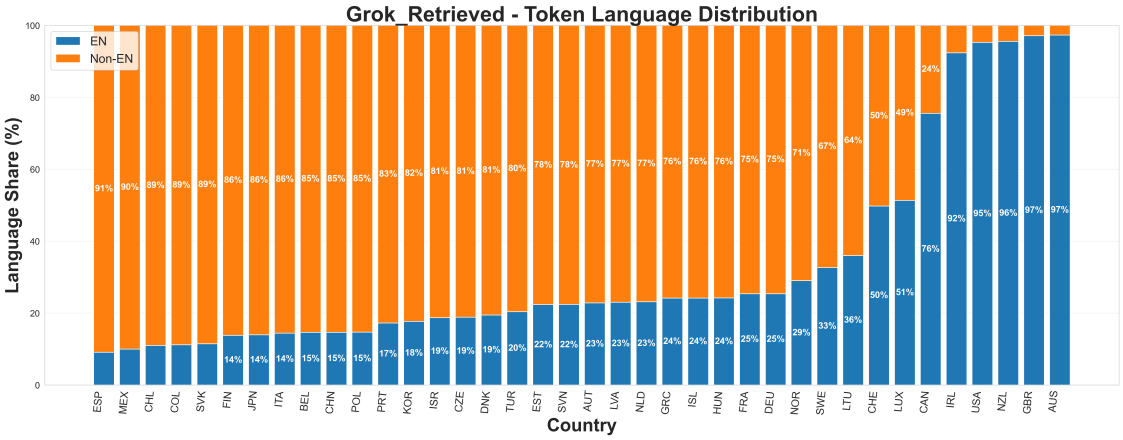

Figure A4.5: Cross-national heterogeneity in retrieved corpora language composition (descriptive analysis). Each bar represents the percentage of non-English resources in the retrieved corpus for each country, illustrating substantial variation in language diversity across the OECD sample. Higher non-English shares are associated with lower extraction performance.

A4.5 Disaggregated mechanism plots

The following plots provide granular diagnostics on recall as a function of context length and language. Figure A4.6 shows the language composition effects on recall with country fixed effects,

demonstrating that higher non-English token shares are associated with lower recall. Figure A4.7 shows the token length effects on recall accross models , demonstrating that higher token length are associated with lower recall.

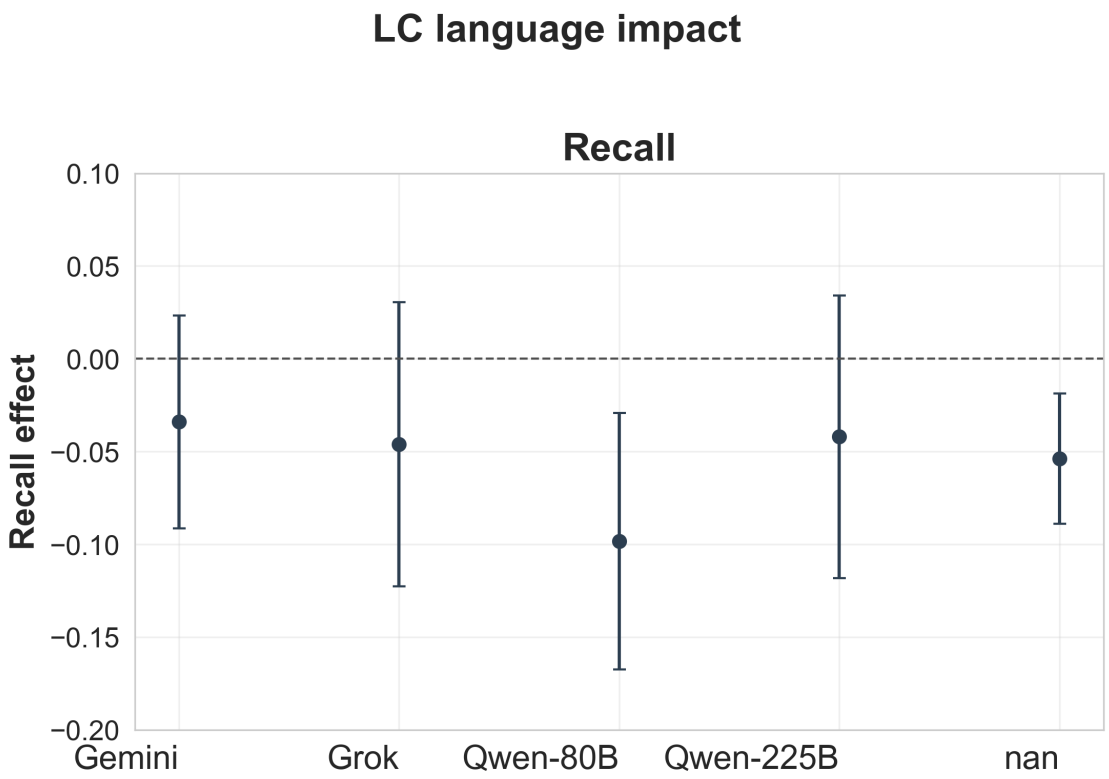

Figure A4.6: Language composition effects on recall (country fixed effects). Higher non-English token shares are associated with lower recall, indicating a quality channel in cross-national retrieval.

A5 Case Study: Erik Solheim Agent Run

This case study details the complete agent execution process for retrieving and synthesizing information about Erik Solheim, former Norwegian Minister of the Environment (2007–2012) and Executive Director of UN Environment Programme (2016–2018). We illustrate how the Supervisor–Searcher architecture (Section A2) operates in practice for a Non-US political figure requiring multilingual evidence synthesis.

A5.1 Agent Execution Overview

The agent processed this case through **31 API calls** across 3 systematic batches, conducting **12 web searches** that returned 156 results, from which 14 documents were retrieved and archived. Table A5.1 summarizes the key metrics.

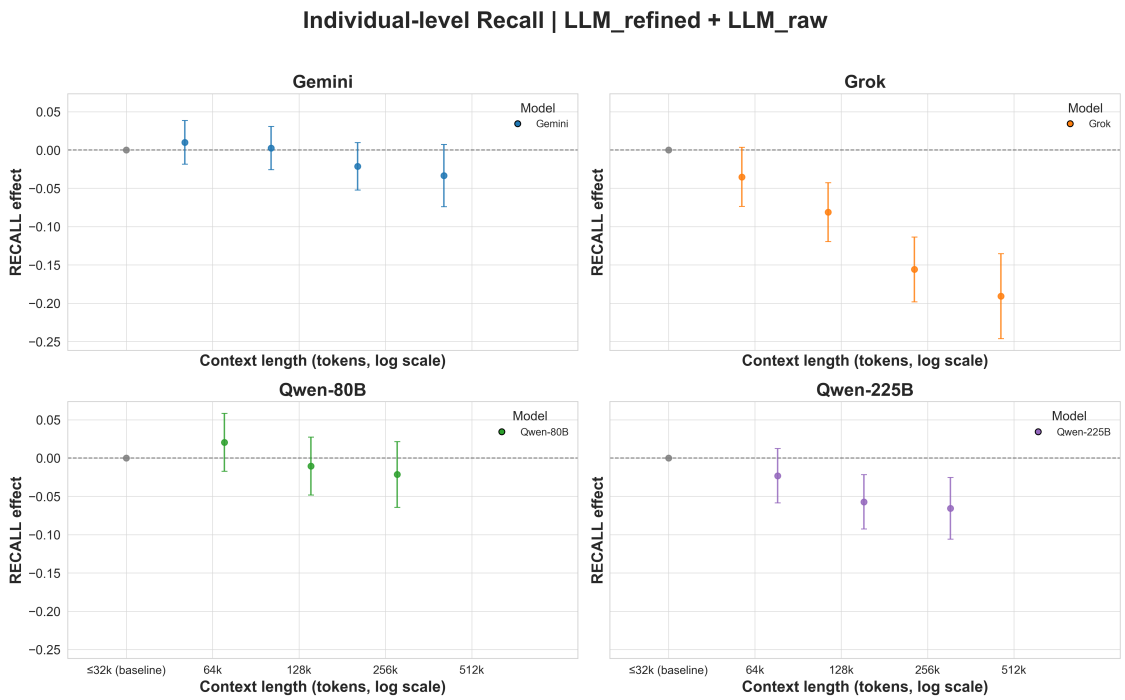

Figure A4.7: Binned recall by context length, disaggregated by model.

A5.2 Three-Phase Search Strategy

The agent execution followed an iterative refinement pattern, transitioning from broad information gathering to targeted gap-filling:

A5.2.1 Phase 1: Initial Skeleton Construction (Messages 0–7)

Supervisor's Goal: Create a comprehensive initial sweep prioritizing official Norwegian government sources and Wikipedia to establish baseline biographical details.

Searcher Execution:

- **Query:** `Erik Solheim AND (biografi OR miljøminister OR SV OR født)`
- **Target Sites:** `no.wikipedia.org`, `regjeringen.no`, `stortinget.no`
- **Language:** Norwegian (NO) and English (EN)

Key Evidence Retrieved:

- *no.wikipedia.org*: "Erik Solheim (født 18. januar 1955)... Sjef for FNs miljøprogram... SV... MDG (2019–)"
- *regjeringen.no (Government CV)*: Confirmed unique combined portfolio of Minister of the Environment and International Development; SV leadership 1987–1997
- *stortinget.no (Parliament)*: "Solheim, Erik (1955-)... Foreldre: Bjørn Elling Solheim og Sophie Grindstad... Cand.mag. 1980"

Table A5.1: Erik Solheim Agent Run: Key Metrics

Metric	Value
Total API Calls	31 messages
Total Input Tokens	474,595
Total Output Tokens	35,583
Total Tokens Used	510,178
Web Searches	12 searches (9 rounds)
Search Results Returned	156 results
URLs Retrieved	14 documents
Retrieved Chunks	18 chunks
Processing Batches	3 batches
Final Report Length	4,371 characters

Archive Update: Identity confirmed (born 1955); baseline timeline established for SV leadership (1987–97), ministerial roles (2005–12), and UNEP (2016–18). Gaps identified: spouse names, early schooling details, and 1985–1987 career gap.

A5.2.2 Phase 2: Gap Filling – Family and Mid-Career (Messages 8–20)

Supervisor's Goal: Target Phase 2 expansion for precise details on early education, spouse names, career gaps (specifically 1985–1987 at Norges Handikapforbund), and post-2018 positions.

Searcher Execution:

- **Query 1:** `Erik Solheim AND (kone OR Sissel Bugge OR Gry Ulverud)`
- **Query 2:** `Erik Solheim AND (Norges Handikapforbund OR konsulent)`
- **Query 3:** `Erik Solheim AND (post-2018 OR MDG OR Green Belt)`

Key Evidence Retrieved:

- *snl.no* (Store norske leksikon): "Gift 1) 1989 med Sissel Bugge... Gift 2) 2004 med Gry Høeg Ulverud... Konsulent i Norges Handikapforbund 1985–1987"
- *linkedin.com*: Post-2018 roles including "President Europe-Asia Center (Jul 2023–)" and "Champion Global Renewables Alliance (Nov 2024–)"

Archive Update: Validated marriages (Sissel Bugge 1989–1998, Gry Ulverud 2004–present) and children; confirmed Consultant role at Norges Handikapforbund (1985–87); mapped extensive post-2018 advisory roles.

A5.2.3 Phase 3: Deep Dive – Early Life and Education (Messages 21–40)

Supervisor's Goal: Identify primary and secondary schools attended before Oslo katedralskole; check memoirs or interviews for early life details in the Bislett area.

Searcher Execution:

- **Query 1:** `Erik Solheim AND (skole OR Bolteløkka OR Ila OR Bislett)`
- **Query 2:** `Erik Solheim AND (oppvekst OR barndom OR tidlig liv)`
- **Language:** Norwegian (NO) only

Key Evidence Retrieved:

- *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"); identified primary school *Bolteløkka skole* and lower secondary *Ila skole*. Timeline now continuous from childhood through 2025.

A5.3 Source Diversity and Language Composition

The agent successfully navigated multilingual evidence, retrieving documents across multiple source types:

Table A5.2: Source Distribution for Erik Solheim Retrieval

Source Type	Count	Examples
Government/Official	6	regjeringen.no CVs, stortinget.no
Reference/Encyclopedia	2	no.wikipedia.org, snl.no
News Media	5	vg.no, nettavisen.no
Professional	2	linkedin.com, hydropower.org
Other	1	oslobyleksikon.no, geni.com

The search strategy demonstrates **adaptive multilingual retrieval**: initial queries combined Norwegian terms ("født", "kone", "skole") with English disambiguation, prioritizing high-credibility Norwegian government sources while using English for cross-verification. This reflects the language composition patterns shown in Figure A4.5.

A5.4 Ground Truth Comparison

Table A5.3 presents the consolidated ground truth (CGT) biography entries and their match categories against the agent output. This comparison reveals both the strengths and limitations of the agentic retrieval process.

A5.5 Key Insights and Analysis

A5.5.1 Discovery Successes

The agent demonstrated strong performance on several fronts:

Table A5.3: Case Study (Ground Truth): CGT Entries and Match Categories

Type	CGT Entry	Match
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.	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 | Central Exec.	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 DAC	FULL_MATCH
Career	2016.06–2018.11 | UN 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	FULL_MATCH
Relatives	mother	FULL_MATCH
Relatives	former spouse	FULL_MATCH
Relatives	spouse	FULL_MATCH
Relatives	child	FULL_MATCH

1. **Iterative Refinement:** Successfully transitioned from broad queries ("biografi OR miljøminister") to targeted searches ("Bolteløkka OR Ila"), demonstrating adaptive query reformulation.

2. **Long-Tail Recovery:** Recovered specific primary and secondary school names (Bolteløkka skole, Ila skole) that represent "needle in the haystack" information requiring precise Norwegian-language queries.

3. **Cross-Source Synthesis:** Integrated information across Wikipedia, government CVs, parliamentary records, encyclopedia entries, and contemporary news sources to build a comprehensive timeline.

4. **Recent Activity Tracking:** Successfully identified post-2018 positions including 2024 appointments (Global Renewables Alliance) through LinkedIn and news sources.

A5.5.2 Coverage Limitations

The comparison with CGT reveals systematic gaps:

1. **Weakly Connected Nodes:** Several concurrent advisory roles (Global Solar Council, Global Wind Energy Council, International Hydropower Association) were missed, suggesting the agent did not exhaustively traverse all post-2018 organizational affiliations.

2. **Minor Positions:** Shorter-term roles (Norwegian Air Force conscript 1974–75, Special Adviser 2012–13, Plastic REVolution Foundation CEO) were not discovered, indicating challenges with brief or less-documented career phases.

3. **Granularity Gaps:** Party membership continuity (1989–2019) was captured as a consolidated period rather than the granular breakdown in CGT, reflecting codebook representation choices.

A5.5.3 Efficiency Analysis

Token usage breakdown reveals the cost structure of agentic synthesis:

Table A5.4: Token Usage Breakdown by Component

Component	Input	Output	Total
Searcher Agent	421,483	32,723	454,206
Coder Agent	53,112	2,860	55,972
Total	**474,595**	**35,583**	**510,178**

The Searcher consumed 89% of total tokens, reflecting the computational cost of processing retrieved documents. The average of 14,534 input tokens per searcher call indicates substantial context accumulation across the multi-turn conversation.

A6 Prompts

We list the main prompt templates used in our pipeline.

```
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}.
```

A6.1 Searcher prompt

```
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}.
```

```
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 words) integrating all
    details.
- Include the source of the information, credible or not, ensure
    reproducibility.
```

A7 A Practical Guide to Information Extraction with LLMs

This appendix provides a practical guide for applying Large Language Models (LLMs) to information extraction tasks in the social sciences. The guiding principle is to treat *extraction* as an end-to-end data-production task and, when necessary, to separate it into two modular stages: **synthesis** (evidence acquisition and refinement) and **coding** (mapping a refined corpus into a structured codebook). This modular design improves auditability and helps us diagnose whether errors arise from missing evidence (a synthesis failure) or incorrect mapping (a coding failure).

A7.1 A minimal workflow for reliable extraction

Step 0 (define the record and the evidence rule). Extraction is only well-defined relative to a codebook. Before using an LLM, we specify (i) a field-level codebook (variables, types, allowed formats), (ii) normalization rules (names, organizations, dates), and (iii) a groundedness requirement (what constitutes sufficient evidence for a claim). In political fact extraction, a small number of ambiguous fields (e.g., office titles, start/end dates) can drive large downstream measurement error, so explicit formatting and disambiguation rules are essential.

Step 1 (diagnose whether synthesis is necessary). The most important practical decision is whether the available sources are effectively **curated** or **open-ended and noisy**. When a short, high-signal source exists (e.g., Wikipedia, an official CV, a curated archive), we can often run **coding-only**: we provide the curated text and ask the model to output the structured record.[23] When relevant evidence is dispersed across many documents (e.g., the open web) or the total context exceeds any fixed window, **synthesis is necessary**: we must decide which sources to read and how to condense them into a signal-dense representation before coding can be valid.

Step 2 (implement coding with constraints and groundedness). The coding stage maps a fixed input corpus into a structured record. In practice, we recommend three safeguards: constrain outputs to be strictly codebook-conformant (e.g., JSON with fixed keys and date formats); require evidence pointers (quotes/snippets) for each claim to reduce hallucination risk (Mallen et al. 2023);

[23] Even in curated settings, long contexts can degrade reliability when relevant facts are buried deep in lengthy inputs (Liu et al. 2024).

and, when feasible, separate generation from validation (a second pass or second model that checks codebook compliance and evidence support). Modern LLMs can often perform this stage in a zero-shot or few-shot manner when the input is curated and the codebook is explicit (Ornstein et al. 2025; Ziems et al. 2024).

Step 3 (implement synthesis as bounded, credibility-aware refinement). Synthesis is an evidence-refinement process: retrieve, filter, cross-check, and compress information into a corpus that is feasible for coding. While there are multiple valid implementations (keyword search, embedding retrieval, human-in-the-loop), open-ended political fact extraction often requires an **interactive** process because early discoveries change what should be searched next. Agentic workflows operationalize this by alternating between reasoning and tool use (ReAct) (Yao et al. 2023), enabling adaptive query refinement and iterative evidence accumulation. Operationally, we recommend explicit retrieval budgets (steps/tokens/sources), credibility-aware filtering (prioritize authoritative sources; deduplicate near-identical content), and compression with traceability (store a refined corpus plus source-linked snippets so claims remain auditable).

Step 4 (evaluate and interpret trade-offs). For extraction, both false positives and false negatives are substantively costly. Precision captures whether extracted claims are correct; recall captures whether the system recovers relevant claims; and F1 summarizes the trade-off. When precision is low, the coding stage is often hallucinating or mis-mapping (tighten groundedness and codebook constraints; improve synthesis filtering). When recall is low, the system is missing evidence (increase synthesis coverage or improve the refined representation).

References

Exa (Nov. 2025). *New Deep Search Type*. Exa Search API changelog entry. Exa. URL: https://exa.ai/docs/changelog/new-deep-search-type (visited on 02/2026).

Liu, N. F., K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang (2024). "Lost in the middle: How language models use long contexts." In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173.

Mallen, A., A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi (2023). "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott (2025). "How to train your stochastic parrot: Large language models for political texts." In: *Political Science Research and Methods* 13.2, pp. 264–281.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao (2023). "ReAct: Synergizing reasoning and acting in language models." In: *International Conference on Learning Representations*.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang (2024). "Can large language models transform computational social science?" In: *Computational Linguistics* 50.1, pp. 237–291.