LLMpedia: A Transparent Framework to Materialize an LLM's Encyclopedic Knowledge at Scale

Anonymous ACL submission

Abstract

Benchmarks such as MMLU suggest flagship language models approach factuality saturation, with scores above 90%. We show this picture is incomplete. *LLMpedia* generates encyclopedic articles entirely from parametric memory, producing ∼1M articles across three model families without retrieval. For gpt-5-mini, the verifiable true rate on Wikipedia-covered subjects is only 74.7% – more than 15 percentage points below the benchmark-based picture, consistent with the availability bias of fixed-question evaluation. Beyond Wikipedia, frontier subjects verifiable only through curated web evidence fall further to 63.2% true rate. Wikipedia covers just 61% of surfaced subjects, and three model families overlap by only 7.3% in subject choice. In a capture-trap benchmark inspired by prior analysis of Grokipedia, LLMpedia achieves substantially higher factuality at roughly half the textual similarity to Wikipedia. Unlike Grokipedia, every prompt, artifact, and evaluation verdict is publicly released, making LLMpedia the first fully open parametric encyclopedia – bridging factuality evaluation and knowledge materialization. All data, code, and a browsable interface are at https://llmpedia.net.

1 Introduction

Large language models are now used as everyday knowledge interfaces by hundreds of millions of people (OpenAI, 2025b). Yet what these models actually *know* – and how reliably they express it in long-form text – remains poorly understood. Two research threads address this from opposite directions: *factuality evaluation* probes models on experimenter-chosen questions; *knowledge materialization* lets models produce their own knowledge at scale. Neither alone suffices: benchmarks miss what they never ask, and materializations lack systematic evaluation. LLMpedia bridges the two.

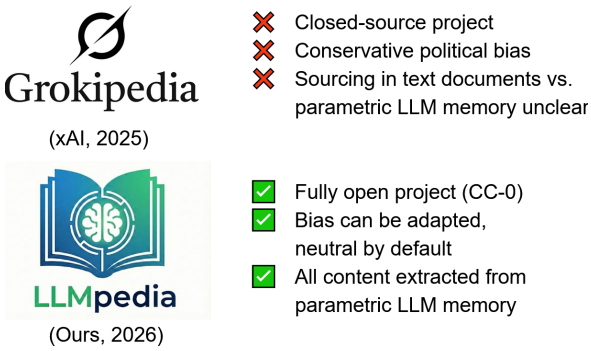

Figure 1: LLMpedia generates entirely from parametric memory with full auditability. Grokipedia's opaque pipeline shows evidence of retrieval-shaped generation (Yasseri, 2025).

Most current evaluation relies on curated benchmarks such as MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2022), and HLE (Phan et al., 2024). These are indispensable, but they share a structural limitation: they test only knowledge the evaluator thought to ask about. This creates a form of *availability bias* (Tversky and Kahneman, 1973): a model can look highly factual on a fixed set of questions while containing broad regions of weak or unstable knowledge the benchmark never reaches. The limitation becomes sharper in real use, where people receive fluent, multi-paragraph explanations rather than short benchmark answers (Saeed et al., 2025a).

A natural alternative is to *materialize* model knowledge at scale. Cohen et al. (2023) and later GPTKB (Hu et al., 2025) showed that recursive elicitation can surface vast amounts of model knowledge, including large regions absent from Wikipedia (Hu et al., 2026; Ghosh et al., 2025). But GPTKB materializes triples – a format unlike how users actually interact with models (OpenAI, 2022) – and its pipeline suffers from persistent entity disambiguation failures (Ghosh et al., 2025; Hu et al., 2026). We instead materialize full en-

cyclopedic articles in the free-text modality users see, making it possible to evaluate subject choice, discourse structure, entity salience, and – crucially – how much apparent knowledge survives outside the narrow corridor of benchmark questions.

This gap is large, for gpt-5-mini, the verifiable true rate on Wikipedia-covered subjects is only 74.7% – more than 15 percentage points below the 90%+ picture suggested by benchmark-style evaluation. That gap is the main empirical signal of this paper: fixed-question benchmarks overstate open-ended factual reliability *in part* because they are shaped by availability bias in what gets asked. Among surfaced subjects absent from Wikipedia but recoverable through curated web evidence, the true rate falls further to 63.2%. At the same time, Wikipedia covers only 61% of surfaced subjects, and three model families overlap on just 7.3%, indicating that materialized parametric knowledge is far more model-specific than leaderboard-style evaluation implies.

This question matters beyond benchmarking. Wikipedia's coverage is uneven, with well-documented Western-centric bias (Wikipedia contributors, 2025b), gender skew (Wikipedia contributors, 2025a), and non-Anglophone gaps. If models encode reliable knowledge beyond Wikipedia, materializing it could help complement those gaps – but only if the generation process is auditable. The most visible LLM-generated encyclopedia, Grokipedia (xAI, 2025), is methodologically opaque. Prior work reports high textual similarity to Wikipedia (Yasseri, 2025), and we document additional evidence consistent with retrieval-shaped generation (Appendix K). This creates a *capture trap*: an encyclopedia may inherit Wikipedia's phrasing while obscuring both genuine model knowledge and its error modes.

We present *LLMpedia*, a fully auditable, open-source framework for parametric encyclopedia generation. Starting from a seed, the system expands breadth-first through article generation and wikilink extraction, then filters candidates through a three-stage sanitization pipeline before re-enqueuing them. Every prompt, intermediate artifact, and evaluation verdict is publicly released. We operate in two modes: *general-domain materialization*, where BFS expands freely from a seed entity, and *topic-focused generation*, where expansion is constrained to entities associated with one of three root topics. Both regimes use explicit prompt templates documented in Appendix J.

System	Param.	Fact. Full Art.	Serendip.	Scale
Benchmarks				
MMLU (Hendrycks et al., 2020)	✓	✗	✗	10^4
HLE (Phan et al., 2025)	✓	✗	✗	10^3
Knowledge materialization				
GPTKB (Hu et al., 2025)	✓	✗	✓	10^6
Cosmopedia (Allal et al., 2024)	✓	✗	(✓)	10^7
Encyclopedia generation				
WikiSum (Liu et al., 2018)	✗	✓	✗	10^4
STORM (Shao et al., 2024)	✗	✓	✗	10^2
Grokipedia (xAI, 2025)	✗(?)	✓	?	10^6
LLMpedia (ours)	✓	✓	✓	10^6

Table 1: Comparison across paradigms: benchmark-based probing, large-scale knowledge materialization, and encyclopedia generation. Param.=parametric-only generation (no retrieval at inference time), Fact. Full Art.=generates factual full text, Serendip.=enables not preconceived knowledge discovery.

Across gpt-5-mini (OpenAI, 2025a), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and DeepSeek-V3 (Liu et al., 2024), we produce approximately one million articles. Motivated by claims of ideological bias in Grokipedia, we additionally inject controlled editorial personas (scientific-neutral, left-leaning, conservative) across all pipeline stages in topic-focused runs, treating framing as an explicit, auditable variable rather than a hidden confounder. Our investigation centers on three questions:

RQ1: What do topic-focused parametric encyclopedias reveal about cross-model factuality, coverage, and the effect of editorial persona?

RQ2: At 10^5 scale, how do three model families differ in knowledge breadth, overlap, and reliability? Does factuality degrade with BFS depth?

RQ3: Does parametric generation escape the Wikipedia capture trap – producing factually competitive but structurally independent content?

Our contributions are:

1. A scalable, fully open-source pipeline for parametric encyclopedia construction – bridging factuality evaluation and knowledge materialization into a single framework;
2. Evidence that open-ended parametric generation is substantially weaker than benchmark-style factuality suggests: for gpt-5-mini, true rate falls to 74.7% on Wikipedia-covered subjects, more than 15 points below the benchmark-based picture;
3. Evidence that the beyond-Wikipedia frontier is both real and difficult: subjects verifiable only through curated web evidence reach just 63.2% true rate;

4. A capture-trap benchmark, inspired by Yasseri (2025), in which we collect the most-edited English Wikipedia articles and show that LLMpedia achieves higher factuality than Grokipedia at roughly half the textual similarity to Wikipedia.

2 Related Work

Knowledge probing and its limits. Probing frameworks such as LAMA (Petroni et al., 2019) test recall via cloze prompts; multitask suites like MMLU (Hendrycks et al., 2020) span knowledge-heavy domains; TruthfulQA (Lin et al., 2022) targets misconceptions; HLE (Phan et al., 2025) pushes frontier difficulty. These measure only what the experimenter thought to ask. LLMpedia complements benchmarks by materializing knowledge at scale and decomposing outputs into true, false, and unverifiable.

From triples to articles. Cohen et al. (2023) introduced recursive elicitation via knowledge-graph expansion; Hu et al. (2025) scaled this into GPTKB (100M+ triples), revealing that many materialized entities are absent from Wikipedia (Ghosh et al., 2025). GPTKB materializes triples; LLMpedia materializes full discourse-level articles. This is not merely a length difference: articles make it possible to evaluate structure, subject choice, and unverifiability in free text.

Encyclopedia generation and the retrieval trap. Early work assumed structured inputs (Sauper and Barzilay, 2009; Lebret et al., 2016); WikiSum (Liu et al., 2018) scaled to multi-document summarization. STORM (Shao et al., 2024) models pre-writing through web-grounded interactions but remains retrieval-dependent. Gao et al. (2024) found that outlines improve generation; LLMpedia extends this with dynamic outline generation tailored to each entity. Grokipedia (xAI, 2025) operates at the largest scale but discloses no methodology; Yasseri (2025) report TF-IDF cosine of 0.455–0.489 against Wikipedia, nearly double LLMpedia's 0.256, suggesting retrieval-shaped repackaging.

Hallucination, retrieval, and provenance. Hallucination remains a central problem for factual generation (Ji et al., 2023). Retrieval-augmented generation (RAG) can improve factuality by grounding outputs in external sources (Lewis et al., 2020; Mallen et al., 2023), while self-reflective variants and confidence elicitation can further improve reliability (He and Razniewski, 2025; Tian et al., 2023). But retrieval complicates provenance: once external passages are introduced, it becomes hard to separate what comes from parametric memory from what is copied or assembled from retrieved text. In encyclopedia-style generation, this matters because retrieval can pull systems toward sources such as Wikipedia, obscuring whether the system is expressing model knowledge or producing retrieval-shaped content (Tuquero and PolitiFact, 2025; Yasseri, 2025).

3 LLMpedia Framework

Figure 2 shows the full pipeline. From a seed entity, the system expands breadth-first through optional self-grounding, outline generation, article elicitation, and three-stage entity sanitization before enqueuing surviving children. All prompts and intermediate artifacts are logged. Full prompt templates for both general-domain and topic-focused runs appear in Appendix J.

3.1 Materialization Modes

General-domain expansion. In the general-domain setting, BFS starts from a single seed entity – *Vannevar Bush* in all reported runs – and expands without topical restrictions. Each article yields Wikitext with `[[wikilinks]]`; after sanitization, surviving links become the next frontier. This mode surfaces an unrestricted knowledge graph and scales to 10^5–10^6 articles.

Topic-focused generation. The topic-focused setting uses the same recursive BFS logic, but constrains expansion around one of three root topics: *Ancient Babylon*, *US Civil Rights Movement*, and *Dutch Colonization in Southeast Asia*. We seed each run with a topic-specific subject list and instruct the model to generate articles whose `[[wikilinks]]` remain directly associated with the root topic. Only topically relevant survivors are re-enqueued as children. This yields a controlled version of BFS useful for cross-model comparison, persona analysis, and domain-difficulty analysis. Prompts for both regimes are documented in Appendix J.

3.2 Self-Grounding

Before article generation, the model can optionally produce a structured fact sheet – summary,

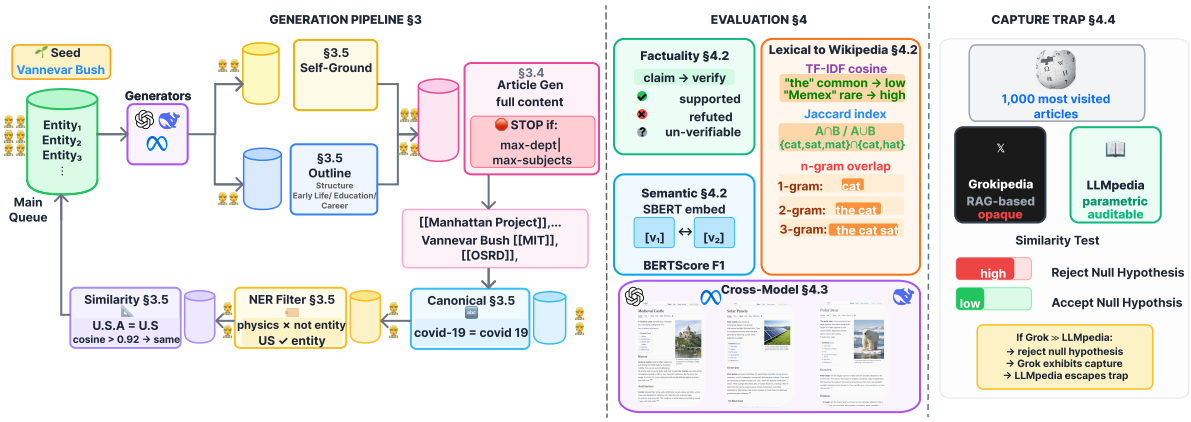

Figure 2: LLMpedia pipeline. Each subject flows through optional self-grounding, dynamic outline generation, and article elicitation. Extracted `[[wikilinks]]` undergo canonical normalization, LLM-based encyclopedic filtering, and embedding-based deduplication before surviving children enter the BFS queue. Details in Appendix E.

```
Subject:  Vannevar Bush
Outline:
1.  Early life and education
2.  Engineering career and Raytheon
3.  Wartime science leadership
...
```

Figure 3: Dynamic outline for *Vannevar Bush*. Sections adapt to the entity rather than following a fixed template.

```
'''Vannevar Bush''' (March 11, 1890 -- June 28,
1974) was an American [[engineer]], [[inventor]],
and [[science administrator]] who headed the U.S.
[[Office of Scientific Research and Development]]
during [[World War II]]. ...

== Early life ==
Born in [[Everett, Massachusetts]], Bush studied
at [[Tufts University]] ...

[[Category:American engineers]]
```

Figure 4: Generated Wikitext excerpt. Blue: retained entities for expansion; red: filtered by Stage 2 (§3.5) – common nouns and generic roles that would produce circular, low-quality sub-articles.

aliases, and predicate–object pairs with calibrated confidence scores – to anchor downstream content and reduce cross-section drift. Facts below threshold 0.75 are flagged. We treat self-grounding as an ablation variable (§4.4); the final large-scale operating point disables it for simplicity.

3.3 Dynamic Outline and Article Generation

Prior work found that outlines improve LLM-based encyclopedia generation (Gao et al., 2024), but mainly studied fixed section templates. LLMpedia instead uses dynamic outline generation: the model first proposes 6–7 section headings tailored to the entity type, then writes the article to match that outline. If warranted, the model also produces an `{{Infobox}}` block.

The model then generates Wikitext following this outline: bolded title, concise lead, `[[wikilinks]]`, and optional `[[Category:...]]` tags. A typical article yields dozens of raw wikilink candidates, motivating the sanitization pipeline below.

3.4 Prompting Strategies

The *baseline* prompt requests standard Wikitext with plain `[[wikilinks]]`. The *calibrated* prompt instead requires per-entity confidence annotations (`[[Entity (0.85)]]`); entities below threshold 0.75 are discarded before NER, making coverage–factuality tradeoffs explicit (Tian et al., 2023). The exact prompt families for general-domain and topic-focused runs are given in Appendix J.

3.5 Entity Sanitization Pipeline

Both GPTKB (Hu et al., 2025) and GPTKB v1.5 (Hu et al., 2026) document persistent entity disambiguation and fragmentation failures. We address these with three stages; Figure 5 illustrates the effect on the general-domain expansion graph, where only encyclopedic entities survive to spawn sub-graphs.

Stage 1: Canonical Normalization. Unicode NFKC, case-folding, and punctuation removal map surface variants to identical keys (e.g., "John F. Kennedy" → `john f kennedy`). Fast and deterministic, this eliminates trivial duplicates before costlier stages.

Stage 2: LLM-Based Encyclopedic Filtering. Not every wikilink warrants its own article. LLMs exhibit a *looping problem*: expanding generic terms like `[[engineer]]` or `[[inventor]]` (Figure 4) yields thin articles that in turn produce equally generic wikilinks, diluting quality at each expansion layer. In contrast, `[[Office of Scientific Research and Development]]` expands into a rich, self-contained article with distinctive downstream entities. Traditional NER tools degrade on isolated entity labels; following GPTKB (Hu et al., 2025), we use the LLM itself to classify candidate batches (prompt in Appendix J), retaining entities that genuinely deserve standalone encyclopedia entries.

Stage 3: Embedding-Based Deduplication. Canonical normalization misses semantic duplicates ("Deutschland" vs. "Germany"). We maintain an embedding index (`text-embedding-3-small`) over all committed entities; candidates exceeding cosine threshold 0.90 are flagged and resolved via LLM arbitration with parent-article context. In parallel execution, concurrent NER workers may independently nominate the same candidate from different parents within the same BFS wave; within-wave cross-candidate deduplication at the similarity worker collapses these before any commit, guaranteeing the main queue contains no duplicates (Appendix E.2).

3.6 Architecture and Persona Injection

Each pipeline stage maintains an independent work queue with retries and exponential backoff, supporting both batch and online modes (Appendix E.1). To study framing as an explicit variable, we inject three personas – scientific-neutral, left-leaning, and conservative – at the system level of *all* pipeline stages in topic-focused runs. This changes not only article prose, but also outline generation, link proposal, and filtering behavior.

4 Experimental Setup

4.1 Models

We evaluate gpt-5-mini (OpenAI, 2025a), Llama-3.3-70B-Instruct (Grattafiori et al., 2024) (4×A100), and DeepSeek-V3-0324 (671B total / 37B active MoE) (Liu et al., 2024). All runs use temperature 0 with fixed seeds. gpt-5-mini serves as the primary model for its instruction-following compliance and cost efficiency; the open-weight models test generalization beyond a single API system. Further implementation and cost details appear in Appendix B.

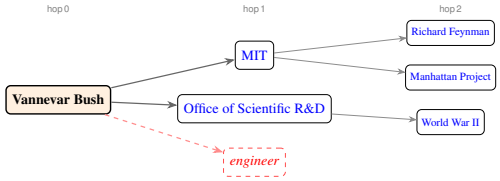

Figure 5: BFS expansion from seed *Vannevar Bush*. Blue nodes survive sanitization and enter the queue; red dashed nodes are filtered as generic.

4.2 Experimental Conditions

Topic-focused generation (RQ1). Three culturally diverse root topics – Ancient Babylon, US Civil Rights Movement, and Dutch Colonization in Southeast Asia – each seed approximately 1K subjects under the topic-focused prompt regime. Each topic is generated across all three models and all three editorial personas (27 configurations, ∼27K articles).

Large-scale materialization (RQ2). All three models start from the same seed entity (*Vannevar Bush*) under general-domain BFS expansion. Recursive BFS produces approximately 741K articles for gpt-5-mini and approximately 120K each for the open-weight models; we compare all three on their first 120K inserted subjects.

Capture-trap benchmark (RQ3). Following Yasseri (2025), we collected the 1,000 most-edited English Wikipedia articles and confirmed their presence on Grokipedia. We then instructed LLM-pedia to generate exactly these 1,000 subjects, targeting Wikipedia-like structural statistics. Both systems are evaluated against Wikipedia *and* curated web sources.

4.3 Evaluation Protocol

Our automatic verification pipeline has two steps. First, an LLM judge (`gpt-4.1-nano`) extracts up to ten atomic, verifiable factual claims per article. Second, each claim is checked against reference evidence and classified as *supported*, *refuted*, or *insufficient evidence*.

Crucially, factuality metrics are computed only over subjects for which reference evidence exists. For **Tier 1**, the full Wikipedia article is retrieved via the MediaWiki API with redirect resolution.

SG	Prompt	Reas.	Prec.%	True%	False%	Unv.%
✓	calib	min	**96.3**	**93.7**	3.4	2.9
✓	base	min	96.2	93.0	3.6	3.4
✗	calib	min	94.5	89.7	4.5	5.8
✗	base	min	94.7†	89.5	4.5	6.0

Table 2: Ablation on gpt-5-mini (minimal reasoning, 500-article sample). †: selected operating point. Full 2^3 table in Appendix A.

Subjects absent from Wikipedia contribute to coverage statistics but not to precision, true rate, or false rate. For **Tier 2**, claims about frontier subjects absent from Wikipedia are checked against curated web sources drawn from a quality-scored domain list (133 domains; Appendix D). Because this evidence pipeline is intentionally strict, some Wikipedia-absent subjects still yield no usable web evidence and are excluded from Tier 2 factuality metrics. In the capture-trap experiment (§5.4), both tiers run for all subjects.

Metrics. Let N be the total number of extracted claims for an article, with n_s supported, n_r refuted, and n_u insufficient:

$$\text{True rate} = n_s/N \quad (1)$$

$$\text{False rate} = n_r/N \quad (2)$$

$$\text{Unverifiable rate} = n_u/N \quad (3)$$

$$\text{Precision} = n_s/(n_s + n_r) \quad (4)$$

The first three partition every claim and sum to 1. Precision excludes unverifiable claims from the denominator. All factuality metrics are macro-averaged at the article level unless otherwise noted.

Similarity. We compute TF-IDF cosine, token Jaccard, n-gram overlap (n=1–3), and semantic cosine (`text-embedding-3-small`) between generated articles and Wikipedia.

4.4 Ablation

We conduct a 2×2×2 factorial ablation over **self-grounding** (on/off), **prompting strategy** (baseline vs. calibrated), and **reasoning budget** (minimal vs. low) on gpt-5-mini. For each of the eight configurations, we generate 5K articles and evaluate a fixed random sample of 500 articles (seed 42) against full Wikipedia pages using `gpt-4.1-nano` as judge. Table 2 reports the four minimal-reasoning configurations; the full 2^3 design appears in Appendix A.

The best configuration (**SG + calibrated + minimal**) reaches 96.3% precision and 93.7% true rate, but self-grounding adds approximately 33% cost and additional pipeline complexity. We therefore adopt **no SG + baseline + minimal**† as the default operating point for all large-scale runs, prioritizing simplicity and reproducibility over the last few points of peak performance.

5 Results

5.1 Entity Sanitization at Scale

The main scaling challenge is not article generation alone, but sanitizing the much larger expansion frontier that generated articles create. GPT-5-mini produces articles for about 740K subjects, while Llama and DeepSeek produce articles for about 120K subjects each. Yet even these smaller open-model runs emit very large numbers of candidate expansion links: 51.6M raw wikilink candidates for GPT-5-mini, 9.8M for Llama, and 6.5M for DeepSeek.

Most of these candidates do not become new subjects. They are removed in three stages: lexical/canonical deduplication, NER filtering, and semantic deduplication. After these stages, 741K candidates survive as new queued subjects for GPT-5-mini, compared with 498K for Llama and 396K for DeepSeek. These final counts are frontier nodes admitted into the BFS queue, not completed articles.

The dominant loss comes from redundancy. Relative to raw candidates, only 3.5% of GPT-5-mini candidates survive canonical deduplication, compared with 12.7% for Llama and 13.2% for DeepSeek. Among canonical survivors, NER accepts 71.9% for GPT-5-mini, 54.8% for Llama, and 51.7% for DeepSeek; similarity filtering then retains 60.2%, 73.0%, and 89.3%, respectively. Overall, only 1.44% of GPT-5-mini raw candidates, 5.07% of Llama candidates, and 6.10% of DeepSeek candidates become new queued subjects. This gap reflects the fact that each article emits many wikilinks, many of which are duplicate surface forms, generic non-entities, or semantic near-duplicates.

5.2 RQ1: Topic-Focused Analysis

Each model generates 1K articles per domain for three culturally diverse topics, each under three editorial personas, yielding 27 conditions in total.

	GPT-5-mini	DeepSeek	Llama
Generated articles (subjects)	740K	120K	120K
Raw candidate `[[wikilinks]]`	51.6M	6.5M	9.8M
After canonical dedup	1.81M	859K	1.25M
After NER filtering	1.30M	444K	684K
After similarity filtering	784K	396K	499K
New queued subjects	741K	396K	498K
Raw → queue survival	1.44%	6.10%	5.07%

Table 3: Cross-model sanitization funnel. Generated articles counts completed article-bearing subjects; final counts are new subjects admitted to the BFS queue, not completed articles.

Topic	Model	Prec	Hall	Unv	W%	CaJ
Ancient Babylon	GPT-5-mini	93.8	5.1	8.6	73.8	
	DeepSeek	87.9	9.9	10.0	78.1	.19
	Llama-70B	81.4	15.5	11.5	82.2	
US Civil Rights	GPT-5-mini	93.6	5.6	5.1	87.2	
	DeepSeek	85.4	12.5	7.9	89.4	.22
	Llama-70B	85.0	12.7	9.3	88.9	
Dutch Colon.	GPT-5-mini	93.9	4.9	10.2	75.4	
	DeepSeek	82.6	14.1	11.3	76.1	.17
	Llama-70B	77.6	18.3	12.9	74.4	

Table 4: Topic-focused results (persona-averaged). **Prec**: factual precision. **Hall**: false rate. **Unv**: unverifiable rate. **W%**: Wikipedia coverage. **CaJ**: average canonical entity Jaccard across model pairs for that topic. Factuality is computed over Wikipedia-covered subjects only.

Table 4 reports persona-averaged values; the full per-persona breakdown appears in Appendix H.

GPT-5-mini achieves 93.6–93.9% precision across the three domains; DeepSeek reaches 82.6–87.9%; Llama reaches 77.6–85.0%. The ranking is consistent in all three topics: GPT-5-mini performs best, followed by DeepSeek, then Llama. Dutch Colonization is the hardest topic, with the widest spread between the strongest and weakest model (93.9% vs. 77.6%).

Wikipedia coverage varies substantially by domain: 26.2% of Ancient Babylon, 24.6% of Dutch Colonization, and 12.8% of US Civil Rights subjects have no Wikipedia page. These subjects contribute to coverage statistics but not to Tier 1 factuality metrics. Some Wikipedia-absent subjects also fail to enter Tier 2 because the web-evidence pipeline applies strict source-quality filtering.

Cross-model canonical entity overlap remains low: average canonical entity Jaccard is 0.19 for Ancient Babylon, 0.17 for Dutch Colonization, and 0.22 for US Civil Rights. Thus, even under the same topic constraint, models still foreground different entities. The broad ranking also remains stable under strong prompt-level framing changes.

	GPT-5-mini	DeepSeek	Llama-70B
Corpus properties (120K each)			
Mean `[[wikilinks]]`/art	70.6	51.5	81.0
Mean sections	6.0	5.2	2.1
Mean words	885	739	845
Subject overlap (120K × 3)			
Exact union / inter.	280,494 / 20,417 (7.3%)		
Shared intersection (1K sample, Tier 1)			
Wiki coverage	99.1%	99.1%	99.1%
Precision	**94.2**	87.7	82.4
True rate	89.7	83.6	77.6
Per-model independent (1K sample, Tier 1)			
Wiki coverage	81.5%	81.8%	80.7%
Precision	**91.4**	82.0	75.2
True rate	82.3	75.0	67.0

Table 5: Cross-model comparison at 120K. Factuality is computed over Wikipedia-covered subjects only. Supplementary figures appear in Appendix G.

5.3 RQ2: Cross-Model at 120K

All three models start from the same seed (*Vannevar Bush*) and expand under general-domain BFS. Because wikilink choices are model-dependent, the three graphs diverge rapidly after the first hop. We cap each corpus at 120K subjects and evaluate two 1K samples: one from the *shared intersection* and one from each model's *independent* corpus.

Despite sharing a seed, only 7.3% of subjects appear in all three corpora (20,417 of 280,494 union). Early wikilink choices therefore cascade into substantially different subject graphs. On the shared intersection, GPT-5-mini leads at 94.2% precision, followed by DeepSeek (87.7%) and Llama (82.4%). On independent samples all models decline as coverage falls from 99.1% to roughly 81%. The ranking is stable across both conditions, indicating that the ordering is not merely an artifact of subject selection.

5.3.1 Hop-Stratified and Frontier Verification

We draw a uniform-random sample of 1,000 articles from the full GPT-5-mini corpus and additionally stratify articles by BFS hop. As described in §4.3, factuality metrics are computed only over subjects with available evidence: Wikipedia-covered subjects enter Tier 1; frontier subjects with usable web evidence enter Tier 2.

Table 6 is the backbone of the paper. On the random sample, Wikipedia covers only 613 of 999 returned articles (61.3%). On that covered subset, the true rate is 74.7%, the false rate is 9.8%,

Bucket	Ref.	n	Cov.%	Prec	True	False	Unv
hop 0	Wiki	1	100.0	100.0	100.0	0.0	0.0
hop 1	Wiki	10	100.0	96.0	95.0	4.0	1.0
hop 2	Wiki	184	92.0	93.6	88.2	5.4	6.4
hop 3	Wiki	169	84.5	91.7	84.2	7.0	8.8
hop 4	Wiki	193	64.3	89.8	77.8	7.6	14.6
hop 5+	Wiki	180	60.0	86.6	73.2	8.8	18.1
random	Wiki	613	61.3	87.4	74.7	9.8	15.5
random	Web	767	76.8	81.7	64.9	11.3	23.8
frontier	Web	292	75.7	79.8	63.2	11.9	24.9

Table 6: GPT-5-mini factuality by BFS depth and evidence tier. **Cov.%** is evidence coverage within each bucket: Wikipedia coverage for Tier 1 and usable-web-evidence yield for Tier 2. The frontier row restricts to subjects absent from Wikipedia.

Metric	LLMpedia	Grokipedia
Factuality (Wiki / Web)		
Precision (%)	**89.8 / 87.9**	76.7 / 76.8
True rate (%)	**84.0 / 73.7**	70.0 / 63.2
False rate (%)	**8.4 / 7.2**	20.3 / 15.5
Unverifiable (%)	7.6 / 19.1	9.7 / 21.3
Similarity to Wikipedia		
TF-IDF cosine	0.256	0.492
Bigram overlap	0.089	0.201
Trigram overlap	0.026	0.080
Semantic cosine	0.773	0.811
Mean / Median words	2,016 / 1,958	7,245 / 6,149

Table 7: Capture-trap experiment on 1,000 most-edited Wikipedia articles. Factuality rows report **Tier 1 (Wiki) / Tier 2 (Web)** verdicts. Similarity is computed against Wikipedia only.

and the unverifiable rate is 15.5%. This is the central finding: open-ended materialization reveals a much weaker factual picture than benchmark-style evaluation suggests.

The web tier shows that this gap is not merely a Wikipedia artifact. Across 767 subjects with usable web evidence, precision is 81.7%, true rate 64.9%, false rate 11.3%, and unverifiable rate 23.8%. Restricting further to the frontier subset – subjects absent from Wikipedia – 292 of 386 (75.7%) yield usable web evidence, with 79.8% precision, 63.2% true rate, 11.9% false rate, and 24.9% unverifiable rate.

Across BFS hops, precision declines more slowly than coverage. From hop 2 to hop 5+, Wikipedia coverage falls from 92.0% to 60.0%, while precision falls from 93.6% to 86.6%. What grows most sharply is the unverifiable rate, from 6.4% to 18.1%. The main degradation is therefore not a dramatic surge in refuted claims, but thinning external coverage as BFS moves outward.

5.4 RQ3: Capture Trap

Following (Yasseri, 2025) we collected the 1,000 most-edited English Wikipedia articles confirmed present on Grokipedia. We then instructed LLMpedia to generate exactly these 1,000 subjects while targeting similar structure. Every claim from both systems is verified under two tiers: **Tier 1** against the full Wikipedia article, and **Tier 2** against curated web sources. Similarity metrics are computed against Wikipedia only.

Against Wikipedia, LLMpedia leads Grokipedia by +13.1 pp precision and +14.0 pp true rate. The web tier rules out the interpretation that Grokipedia is merely penalized for correct claims Wikipedia does not mention: the advantage persists at +11.1 pp precision and +10.5 pp true rate, while Grokipedia's false rate remains more than double LLMpedia's under both references.

LLMpedia achieves this at roughly half the TF-IDF cosine (0.256 vs. 0.492) and much lower n-gram overlap, while semantic cosine remains comparable because both systems discuss the same subjects. This is exactly the capture-trap pattern: Grokipedia stays closer to Wikipedia's surface form, yet is less factually reliable.

Despite targeting Wikipedia-like structure, LLMpedia produces shorter articles (2,016 mean words), while Grokipedia overshoots dramatically (7,245). Because our evaluation extracts the same number of claims per article (up to 10), per-claim false rate is not mechanically inflated by length. If Grokipedia's deficit were simply a length artifact, web verification should recover many missed truths by confirming claims Wikipedia omits; instead, its precision stays essentially flat at 76.8%, while LLMpedia drops only modestly from 89.8% to 87.9%. The deficit is systematic.

6 Conclusion

Benchmarks report 90%+ factuality for frontier LLMs, but materialized writing reveals a weaker picture. On a uniform-random GPT-5-mini sample, Wikipedia covers only 61.3% of subjects, with 74.7% true rate on the covered subset – over 15 points below benchmark performance. Web verification confirms the gap: frontier precision is 79.8%. Three model families overlap on just 7.3% of subjects, and on 1,000 shared articles LLMpedia exceeds Grokipedia by +13 pp precision at half the textual similarity to Wikipedia. All data, code, and interface are at `https://llmpedia.net`.

7 Limitations

Single-run conditionality and cost barrier. Every article is the result of a single generation pass under a fixed prompt and temperature-0 decoding. We do not claim that one pass fully characterizes what a model *knows*; it surfaces one sample from a latent knowledge distribution. Self-consistency decoding (Wang et al., 2022) could recover additional knowledge, but the full project already costs approximately $3,000 across generation, ablation, evaluation, and preliminary testing (Appendix B); multi-sample exploration would multiply the dominant generation cost considerably.

Temporal snapshot. All runs were conducted between March and May 2025 with frozen model checkpoints. Wikipedia and web references were retrieved during the evaluation window, but both evolve continuously. A claim judged unverifiable today may become verifiable or refutable later.

Evaluation stack and judge reliability. Automatic LLM-based factuality evaluation is practical at our scale but not perfect. Our pipeline relies on a single judge model, `gpt-4.1-nano`, for both claim extraction and verdict assignment, but this is common practice in research (Hu et al., 2025; Saeed et al., 2025b).

Web evidence coverage and selection bias. For frontier subjects, only 75.7% (292/386) yielded usable web evidence under the strict source-quality filtering described in Appendix D. The remaining 24.3% are excluded from Tier 2 metrics, which may bias frontier estimates toward subjects with stronger public documentation.

Scale asymmetry. Only gpt-5-mini reaches nearly one million articles. The open-weight models are constrained to approximately 120K–498K, so the deepest hop-stratified analysis is GPT-5-mini specific.

Scope of measurement. Our framework captures propositional factual support, but not everything that matters in an encyclopedia article. It does not directly measure coherence, neutrality, salience, completeness, or omission of important facts. An article can therefore achieve high precision while still being editorially unbalanced or incomplete.

Model and data-source limitations. Our study relies on a mixed model stack, including OpenAI API models and open-source models, and the behavior of API-served systems may change over time as providers update them. This limits strict reproducibility across dates and deployments. In addition, although our goal is to analyze encyclopedic generation rather than surface or publish personal data, generated articles and public-source evidence may occasionally mention real individuals because encyclopedic subjects often include public figures. We did not design the pipeline to expose private or sensitive personal information, and any such content that is found inadvertently should be filtered from releases and removed from public-facing interfaces or downloadable artifacts.

References

Loubna Allal, Anton Lozhkov, and Daniel van Strien. 2024. Cosmopedia: how to create large-scale synthetic data for pre-training large language models. Hugging Face Blog.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of EACL*.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Tianwei She, Yuang Jiang, and Irene Li. 2024. Evaluating large language models on Wikipedia-style survey generation. In *Findings of ACL*.

Shrestha Ghosh, Luca Giordano, Yujia Hu, Tuan-Phong Nguyen, and Simon Razniewski. 2025. Mining the mind: What 100m beliefs reveal about frontier llm knowledge. *arXiv*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Jingbo He and Simon Razniewski. 2025. Combining self-retrieval-augmented generation with divide-and-conquer for language model-based knowledge base construction. In *KBC-LM and LM-KBC @ ISWC*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, Moritz Müller, and Simon Razniewski. 2026. Gptkb v1.5: A massive knowledge base for exploring factual llm knowledge. *AAAI*.

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2025. Enabling LLM knowledge analysis via extensive materialization. In *ACL*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *EMNLP*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, NIPS '20.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv:1801.10198*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL*.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/. Accessed: 2026-03-17.

OpenAI. 2025a. GPT-5-mini. Accessed: 2026-03-17.

OpenAI. 2025b. How people are using chatgpt.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv:2501.14249*.

Long Phan and 1 others. 2024. Humanity's last exam. *arXiv*.

Muhammed Saeed, Muhammad Abdul-Mageed, and Shady Shehata. 2025a. Surfacing subtle stereotypes: A multilingual, debate-oriented evaluation of modern llms. *arXiv preprint arXiv:2511.01187*.

Muhammed Saeed, Shaina Raza, Ashmal Vayani, Muhammad Abdul-Mageed, Ali Emami, and Shady Shehata. 2025b. Beyond content: How grammatical gender shapes visual representation in text-to-image models. In *Findings of EMNLP*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *ACL-IJCNLP*.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *NAACL*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*.

Loreben Tuquero and PolitiFact. 2025. What's Grokipedia, Musk's AI-powered rival to Wikipedia? Al Jazeera.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2).

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wikipedia contributors. 2025a. Gender bias on Wikipedia. https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia. [Online; accessed 21-November-2025].

Wikipedia contributors. 2025b. Wikipedia: Systemic bias. https://en.wikipedia.org/wiki/Wikipedia:Systemic_bias. [Online; accessed 21-November-2025].

Wikipedia contributors. 2026. Wikipedia:size of wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=1335945684. Last edited 2026-02-01 02:06 (UTC). Accessed 2026-02-03.

xAI. 2025. Grokipedia. Accessed: 2026-03-17.

Taha Yasseri. 2025. How similar are grokipedia and wikipedia? a multi-dimensional textual and structural comparison. *arXiv:2510.26899*.

A Ablation: Full Results

We conduct a 2×2×2 factorial ablation over **self-grounding** (on/off), **prompting strategy** (baseline vs. calibrated), and **reasoning budget** (minimal vs. low) on `gpt-5-mini`. For each of the eight configurations, we generate 5K articles and evaluate a fixed random sample of 500 articles (seed 42). Claims are verified against full Wikipedia pages by `gpt-4.1-nano`. Table 8 reports all eight cells.

Prompting strategy. Switching from the baseline to the calibrated prompt yields only small and inconsistent gains. Without self-grounding, precision changes from 94.7% to 94.5%; with self-grounding, it changes from 96.2% to 96.3%. The main effect is therefore not a large precision jump, but a modest reshaping of the true–unverifiable balance.

Self-grounding. Self-grounding improves results consistently. For example, under calibrated+minimal, precision rises from 94.5% to 96.3%, and the unverifiable rate falls from 5.8% to 2.9%. Its main benefit is reducing unverifiable claims rather than sharply lowering the false rate. The cost is substantial: approximately 33% higher generation cost, or roughly $850 more per million articles.

Reasoning budget. Increasing the reasoning budget from minimal to low reduces precision in all four self-grounding×prompt settings. The largest drop appears in the no-SG+baseline condition, which falls from 94.7% to 90.7%. The decline mainly comes from a higher false rate, suggesting that longer reasoning traces introduce extra speculative detail rather than improving factual reliability.

Configuration choice. The best configuration is **SG + calibrated + minimal**, which reaches 96.3% precision and 93.7% true rate. We nevertheless choose **no SG + baseline + minimal**† as the default operating point for all large-scale runs. Relative to the best cell, it sacrifices 1.6 pp in precision and 4.2 pp in true rate, but removes self-grounding complexity, uses the simpler baseline prompt, and saves roughly $850 per million articles.

SG	Prompt	Reas.	Prec.%	True%	False%	Unv.%
✗	base	low	90.7	84.6	8.2	7.3
✗	base	min	94.7†	89.5	4.5	6.0
✗	calib	low	92.1	85.4	7.0	7.7
✗	calib	min	94.5	89.7	4.5	5.8
✓	base	low	93.9	89.3	5.5	5.2
✓	base	min	96.2	93.0	3.6	3.4
✓	calib	low	95.3	92.2	4.3	3.6
✓	calib	min	**96.3**	**93.7**	**3.4**	**2.9**

Table 8: Full 2^3 ablation on `gpt-5-mini` (500-article random sample per cell, seed 42). †: selected operating point for all large-scale experiments.

B Models and Costs

B.1 Model Configuration

Table 9 summarizes the models and execution modes used across all experimental conditions. All `gpt-5-mini` runs use the OpenAI Batch API with temperature 0 and minimal reasoning effort. Batch API requests receive a 50% cost reduction relative to standard API pricing; a fraction of NER and similarity calls for the large-scale run were executed via the online API at standard pricing due to batch queue overflow during peak expansion waves.

`Llama-3.3-70B-Instruct` and `DeepSeek-V3-0324` are served on-premise via 4×A100 GPUs through institutional compute infrastructure. Their marginal cost is GPU-hours, which we do not monetize in this analysis.

Embedding-based deduplication uses `text-embedding-3-small` for all models. Automatic evaluation uses `gpt-4.1-nano` as judge for claim extraction and verdict assignment across all experimental conditions.

Model	Execution	Role
`gpt-5-mini`	OpenAI Batch API	generation, NER, similarity
`Llama-3.3-70B`	4×A100 (on-prem)	generation, NER, similarity
`DeepSeek-V3-0324`	4×A100 (on-prem)	generation, NER, similarity
`text-embedding-3-small`	OpenAI Batch API	deduplication embeddings
`gpt-4.1-nano`	OpenAI API	evaluation judge

Table 9: Models used across all experimental conditions.

B.2 Aggregate Cost

The total project expenditure across all API-billed components—including large-scale generation (∼741K articles for `gpt-5-mini`), the full 2^3 ablation study, all 27 topic-focused configurations, the capture-trap benchmark, automatic evaluation of every experimental condition, embedding costs, and preliminary development runs—is

approximately **$3,000**. The open-weight models contribute GPU-hours but no API cost.

C Capture-Trap Benchmark: Full Details

C.1 Experimental Design

The capture-trap benchmark asks whether parametric generation can remain structurally independent on subjects where a retrieval-shaped system has maximal opportunity to track Wikipedia closely.

Title selection. Inspired by Yasseri (2025), we collected 1,000 of the most-edited English Wikipedia titles that were also present on Grokipedia. This deliberately targets high-visibility, heavily maintained subjects where Wikipedia coverage is rich and a retrieval-centered system should have every incentive to converge toward Wikipedia's wording and structure.

Generation. LLMpedia generated all 1,000 articles using `gpt-5-mini` in the baseline large-scale configuration (no self-grounding, minimal reasoning). Grokipedia articles were fetched from the live site. Wikipedia references were retrieved through the MediaWiki API with redirect resolution.

Evaluation. For each title, we extracted up to 10 atomic claims from both systems and verified them against two evidence tiers: **Tier 1** (Wikipedia) and **Tier 2** (curated web sources). Similarity metrics were computed against Wikipedia only. This is the same evaluation pipeline used in the main paper.

C.2 Detailed Results

Table 10 reports the full capture-trap breakdown. Factuality metrics are macro-averaged at the article level and shown as mean ± standard deviation across the 1,000 titles.

Interpretation. The central pattern is an inverse precision–similarity relationship. Grokipedia stays much closer to Wikipedia at the lexical level, with nearly double the TF-IDF cosine and substantially higher n-gram overlap, yet remains markedly less reliable factually. LLMpedia is farther from Wikipedia's surface form while achieving higher precision, higher true rate, and lower false rate under both Wikipedia and web verification. This is the capture trap in practice: lexical proximity to Wikipedia does not imply factual superiority, and may instead reflect retrieval-shaped rewriting without robust disambiguation.

Metric	LLMpedia	Grokipedia
Factuality (Wiki / Web)		
Precision (%)	**89.8±21.2 / 87.9±25.0**	76.7±24.8 / 76.8±25.9
True rate (%)	**84.0±24.6 / 73.7±28.6**	70.0±26.1 / 63.2±27.9
False rate (%)	**8.4±18.0 / 7.2±15.5**	20.3±22.1 / 15.5±16.8
Unverifiable (%)	7.6±15.6 / 19.1±23.5	9.7±13.5 / 21.3±21.2
Similarity to Wikipedia		
TF-IDF cosine	0.256	0.492
Bigram overlap	0.089	0.201
Trigram overlap	0.026	0.080
Semantic cosine	0.773	0.811
Mean / Median words	2,016 / 1,958	7,245 / 6,149

Table 10: Full capture-trap results on 1,000 shared titles. Factuality values are macro-averaged per article and reported as mean ± standard deviation. Similarity is computed against Wikipedia only.

D Web Evidence Pipeline

This appendix documents the Tier 2 verification pipeline for frontier subjects referenced in §4.3 and §5.3.1 (Table 6, frontier row).

D.1 Search and Domain Scoring

The pipeline auto-detects available search backends (Valyu → Serper → Brave → DuckDuckGo). Each retrieved URL receives a quality score on a 0–100 scale, and only sources with score ≥ 60 are eligible for full-page fetching. Tables 12 and 13 list the **133 explicitly scored domains** in the current `evidence.py` implementation. Unknown `.edu`/`.gov`/`.org` domains default to score 70, while unknown HTTPS domains default to 35.

At retrieval time, URLs are first filtered by the blocked-domain list and any user-specified exclusions, then ranked by domain score. The implementation applies exclusions using root-domain matching, so blocking a root domain also excludes its subdomains.

D.2 Blocked Domains

Table 11 lists the blocked root domains used by the current pipeline. These are excluded before evidence retrieval to avoid circular verification, search portals, social media, e-commerce pages, and AI-generated answer engines.

D.3 Content Validation

CAPTCHA challenges, access-denied pages, paywalls, and JavaScript-only responses are detected

Category	Blocked root domains
Circular reference / Wikimedia	wikipedia.org, en.wikipedia.org, en.m.wikipedia.org, wikidata.org, www.wikidata.org, wikimedia.org
Search engines / portals	google.com, bing.com, duckduckgo.com, yahoo.com
Social / community / video	facebook.com, instagram.com, twitter.com, x.com, tiktok.com, reddit.com, quora.com, youtube.com, pinterest.com
E-commerce	amazon.com, ebay.com
AI-generated answer engines	perplexity.ai

Table 11: Root-domain exclusions in the Tier 2 web-evidence pipeline. Exclusion is applied using root-domain matching, so subdomains of blocked roots are also excluded at runtime.

Category	Domain	Score	Domain	Score	Domain	Score
Encyclopedias & reference	britannica.com	100	worldhistory.org	90	encyclopedia.com	85
	scholarpedia.org	88	plato.stanford.edu	95	iep.utm.edu	88
	newworldencyclopedia.org	75				
Government & archives	loc.gov	97	archives.gov	97	congress.gov	95
	usa.gov	93	cia.gov	90	state.gov	90
	whitehouse.gov	90	nasa.gov	95	nih.gov	94
	cdc.gov	93	fda.gov	90	epa.gov	89
	noaa.gov	90	usgs.gov	90	nps.gov	85
	si.edu	92	parliament.uk	88	gov.uk	87
	europarl.europa.eu	85	un.org	88	who.int	90
	worldbank.org	88	imf.org	87		
Wire services & top news	reuters.com	94	apnews.com	94	bbc.com	93
	bbc.co.uk	93	nytimes.com	92	washingtonpost.com	90
	theguardian.com	91	economist.com	90	ft.com	89
	wsj.com	89	theatlantic.com	87	newyorker.com	87
	npr.org	88	pbs.org	88	aljazeera.com	85
	dw.com	84	france24.com	83		
Academic journals & databases	nature.com	95	science.org	95	sciencedirect.com	90
	springer.com	89	link.springer.com	89	wiley.com	88
	onlinelibrary.wiley.com	88	tandfonline.com	87	cell.com	93
	thelancet.com	93	bmj.com	92	nejm.org	94
	pnas.org	92	ncbi.nlm.nih.gov	92	pubmed.ncbi.nlm.nih.gov	92
	jstor.org	90	arxiv.org	85	ssrn.com	82
	researchgate.net	75	scholar.google.com	80	doaj.org	80
	semanticscholar.org	82	ieee.org	89	ieeexplore.ieee.org	89
	acm.org	88	dl.acm.org	88		

Table 12: Explicitly scored domains in the Tier 2 web-evidence pipeline (Part I).

via pattern matching and excluded from caching as valid content. Pages shorter than 200 characters are flagged as unusable. Up to three candidate URLs are attempted in descending quality-score order, with the pipeline falling back to the search snippet when a full-page fetch is blocked.

D.4 Frontier Results Summary

Applied to the 386 Wikipedia-absent subjects in the 1,000-article uniform-random sample, 292 subjects (75.7%) yielded usable web evidence under the domain-quality filtering described above. On this frontier subset, precision is **79.8%**, true rate **63.2%**, false rate **11.9%**, and unverifiable rate **24.9%**. The remaining 94 subjects (24.3%) yielded no usable evidence from the curated web pipeline and are therefore excluded from Tier 2 frontier factuality statistics. This exclusion is discussed in §7 as a source-selection bias: frontier estimates are necessarily conditional on the availability of usable external documentation.

E Pipeline Implementation

E.1 System Configuration and Design Decisions

Execution modes. LLMPedia supports two execution modes that trade latency against cost. In *online mode*, API calls are issued concurrently across a multi-threaded worker pool, with all five pipeline stages running in parallel. A global concurrency cap is enforced across all stages simultaneously to prevent rate-limit exhaustion. In *batch mode*, requests are submitted to the OpenAI Batch API, which processes them asynchronously at a 50% cost reduction with a completion window of up to 24 hours. The two modes are architecturally equivalent and produce identical outputs; batch mode was used in our main experiments for cost efficiency, while online mode is reserved for

Category	Domain	Score	Domain	Score	Domain	Score
Museums, libraries & cultural institutions	smithsonianmag.com	87	nationalgeographic.com	86	metmuseum.org	90
	moma.org	87	nga.gov	88	bl.uk	90
	bnf.fr	88	dpla.org	83	europeana.eu	83
History & biography	history.com	80	biography.com	75	oxforddnb.com	90
	anb.org	88	historytoday.com	78	historyextra.com	77
Science & technology	scientificamerican.com	86	newscientist.com	83	livescience.com	75
	space.com	76	phys.org	78	sciencenews.org	82
	quantamagazine.org	88	arstechnica.com	80	spectrum.ieee.org	84
	technologyreview.com	84				
Education & research organizations	mit.edu	93	stanford.edu	93	harvard.edu	93
	ox.ac.uk	92	cam.ac.uk	92	berkeley.edu	91
	caltech.edu	91	yale.edu	91	princeton.edu	91
	columbia.edu	90	uchicago.edu	90	cornell.edu	89
	cmu.edu	89	ethz.ch	89	mpg.de	89
	khanacademy.org	80				
Fact-checking & data	snopes.com	82	factcheck.org	84	politifact.com	80
	ourworldindata.org	88	statista.com	78	data.gov	88
	census.gov	90	bls.gov	89	bea.gov	88
Legal	law.cornell.edu	90	supremecourt.gov	92	courtlistener.com	80
	oyez.org	82				
Medical & health	mayoclinic.org	88	clevelandclinic.org	85	webmd.com	72
	medlineplus.gov	90	hopkinsmedicine.org	87	uptodate.com	88

Table 13: Explicitly scored domains in the Tier 2 web-evidence pipeline (Part II). Together with Table 12, this yields **133** explicitly scored domains. `scholar.google.com` is explicitly scored in code but still excluded at runtime because `google.com` is blocked by root-domain matching.

small-scale ablations where turnaround time matters.

Model configuration and stage independence. Each pipeline stage—self-grounding, outline generation, article elicitation, NER filtering, and similarity arbitration—can be assigned an independent language model, temperature, and token budget. In practice, all stages within a given run share a single model unless explicitly differentiated. The one principled exception is the embedding model used for similarity deduplication, which is always configured independently: embedding and completion models serve fundamentally different functions and conflating their configuration would silently degrade deduplication quality.

BFS depth and article budget. Expansion depth and total article count are both configurable hard caps. In **topic-focused runs** (§5.2), depth is capped at two hops from the thematic seed to maintain topical coherence. In **general-domain expansion** (§5.3), no depth cap is imposed: GPT-5-mini reaches hop 6 over 741K articles, with 94.2% of the corpus at hop 4 or beyond (Appendix I). In both modes, the seed article at hop 0 is generated but excluded from evaluation metrics to avoid confounding the model's generation behaviour with the trivially known anchor entity.

Similarity threshold and arbitration. The cosine similarity threshold above which two entity embeddings trigger a deduplication check is set to 0.90. This value is conservative: surface-form variants of the same entity (e.g. *United States* and *USA*) consistently fall above it, while genuinely distinct but thematically related entities consistently fall below it. When the threshold is exceeded, a secondary LLM arbitration step is invoked rather than rejecting automatically. This two-stage design prevents both false positives and false negatives.

NER confidence and elicitation filtering. Under the calibrated prompt variants, both the NER stage and the elicitation stage assign explicit confidence scores. Candidates below the NER confidence threshold are rejected before reaching the similarity stage. Rejection at either stage is non-permanent: if a future parent article re-proposes the same entity with sufficient confidence, it re-enters the pipeline from the beginning.

Persona. Each pipeline stage accepts a persona specification injected at the system level. Three personas are provided: scientifically neutral, left-leaning, and conservative. All three share identical structural prompts; only the framing and evaluative language differ.

Fault tolerance and strict-gate semantics. Failed API calls are retried with exponential backoff up to a configurable maximum. After all retries are exhausted, a strict-gate policy applies: a stage that cannot produce a valid output causes the corresponding entity to be *dropped* rather than passed through with a degraded result. For NER, a parse failure produces no candidates rather than passing all candidates unchecked. For similarity arbitration, an API failure causes the candidate to be treated as a duplicate and rejected. This conservative policy introduces a small downward bias on recall but eliminates the risk of contaminating the corpus with candidates that bypassed quality controls.

E.2 Deduplication Correctness Under Parallel Execution

LLMPedia maintains four independent parallel queues: an *elicitation queue*, a *NER queue*, a *similarity queue*, and the *canon queue*. Workers on each queue operate concurrently and independently. This parallelism creates a structural race: multiple NER workers can extract and accept the same entity from different parent articles within the same BFS wave, before any of them reaches the similarity worker. Figure 6 shows a concrete instance and its resolution.

The main article queue is structurally duplicate-free. CANONQUEUE is the *only* channel through which an entity can reach article generation. An entity enters CANONQUEUE if and only if it simultaneously passes Stage 3 similarity *and* is absent from CANONKEYS; both operations execute atomically.

Why commitment happens after similarity, not after NER. A candidate is registered in CANONKEYS only after it passes Stage 3. If commitment happened at NER acceptance, an entity later rejected by similarity would be permanently suppressed even though it was never materialised. Deferring commitment keeps future parent articles free to re-propose the same entity.

Within-batch collisions. A single parent article may produce the same wikilink in two surface forms that canonicalize identically. Within-wave cross-candidate deduplication handles this before any index update.

Correctness guarantee. By induction on BFS waves, the embedding index at the start of wave t contains exactly the entities committed in waves $0, \ldots, t-1$. The main article queue therefore contains no duplicate entities at any point during or after execution.

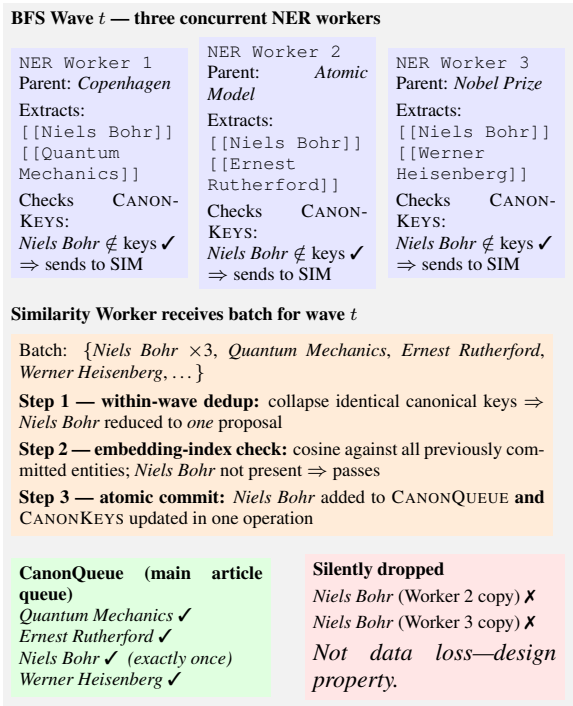

Figure 6: Three concurrent NER workers independently accept [[Niels Bohr]] from different parent articles. Within-wave deduplication collapses the three proposals to one before any commit.

F Entity Sanitization Funnel: Full Breakdown

F.1 Cross-Model Summary

The key distinction in the funnel is between *generated articles* and *new queued subjects*. GPT-5-mini generates articles for about 740K subjects, while Llama and DeepSeek each generate articles for about 120K subjects. Those generated articles then emit much larger expansion frontiers: 51.6M raw candidate wikilinks for GPT-5-mini, 9.8M for Llama, and 6.5M for DeepSeek. After canonical deduplication, NER filtering, and semantic deduplication, the pipeline admits 741K new queued subjects for GPT-5-mini, 498K for Llama, and 396K for DeepSeek. These final counts are frontier nodes added to the BFS queue, not completed articles.

The dominant loss in all three models comes from redundancy. Most emitted wikilinks never become new subjects because they are duplicate surface forms, generic non-entities, or semantic

	GPT-5-mini	DeepSeek	Llama
Generated articles (subjects)	740K	120K	120K
Raw candidate `[[wikilinks]]`	51.6M	6.5M	9.8M
After canonical dedup	1.81M	859K	1.25M
After NER	1.30M	444K	684K
After similarity	784K	396K	499K
New queued subjects	741K	396K	498K
Raw → post-dedup	3.5%	13.2%	12.7%
NER survival	71.9%	51.7%	54.8%
Similarity survival	60.2%	89.3%	73.0%
Raw → queue	1.44%	6.10%	5.07%

Table 14: Cross-model funnel summary. "Generated articles" counts completed article-bearing subjects; "New queued subjects" counts frontier nodes admitted to the BFS queue for future expansion.

near-duplicates of already committed nodes. GPT-5-mini is the most aggressive generator: it emits by far the largest raw frontier, but only 1.44% of raw candidates ultimately become new queued subjects. Llama and DeepSeek generate much smaller article corpora, but each still produces millions of candidate links, of which only 5–6% survive to the queue.

F.2 Survival by BFS Depth

Across models, the main depth effect is a decline in canonical survival, while later-stage filtering is more stable. This pattern indicates graph saturation rather than late-hop collapse in NER or semantic filtering quality: as BFS expands, newly proposed candidates are increasingly redundant with subjects already surfaced earlier.

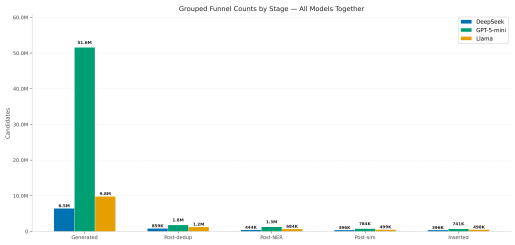

Figure 7: Grouped funnel counts across models. Open models generate only about 120K articles each, but those articles still induce very large candidate frontiers.

F.3 Per-Model Funnels

GPT-5-mini. GPT-5-mini generates articles for about 740K subjects. Those articles emit 51.6M raw candidate wikilinks, which are reduced to 1.81M after canonical deduplication, 1.30M after NER, 784K after similarity filtering, and 741K new queued subjects. Its dominant loss is therefore early redundancy rather than late-stage rejection.

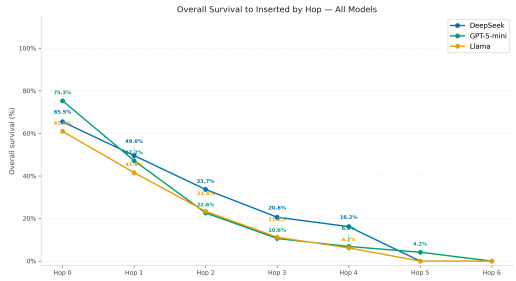

Figure 8: Overall raw-candidate-to-queue survival by BFS hop. The decline reflects increasing redundancy as the queue saturates.

Llama-3.3-70B. Llama generates articles for about 120K subjects, but these still emit 9.8M raw candidate wikilinks. After filtering, 498K new queued subjects survive. A notable artifact is that 270,157 of 563,084 NER rejections are parse failures, so Llama's lower NER survival partly reflects instruction-following brittleness rather than purely semantic filtering.

DeepSeek-V3. DeepSeek also generates articles for about 120K subjects, yielding 6.5M raw candidate wikilinks and 396K new queued subjects after the full funnel. Its similarity survival is the highest of the three models (89.3%), suggesting that once a candidate passes canonical deduplication and NER, it is comparatively less likely to be removed as a semantic duplicate.

G Cross-Model Analysis: Supplementary Figures

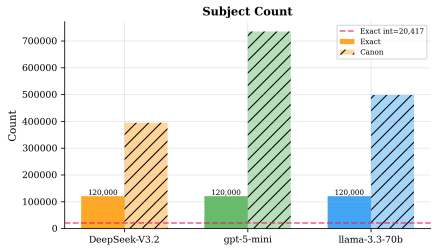

Figure 9: Subject counts with exact three-way intersection. Each model explores ∼120K unique subjects; only 7.3% overlap.

Topic	Model	Persona	n	Prec	Hall	Unv	W%	Wds
A. Bab.	GPT	cons	197	91.8	6.4	9.3	72.4	762
A. Bab.	GPT	left	205	94.0	4.9	8.9	74.3	788
A. Bab.	GPT	neut	202	95.4	4.0	7.4	74.8	770
A. Bab.	DS	cons	216	90.0	8.4	8.3	81.5	925
A. Bab.	DS	left	221	89.8	8.5	9.2	83.4	920
A. Bab.	DS	neut	188	83.2	13.2	12.9	69.6	599
A. Bab.	Llama	cons	217	79.7	16.3	13.2	80.4	967
A. Bab.	Llama	left	201	83.0	14.5	11.0	75.8	983
A. Bab.	Llama	neut	244	81.6	15.5	10.3	90.4	964
D. Col.	GPT	cons	202	95.2	4.0	9.5	74.8	819
D. Col.	GPT	left	211	94.3	4.8	10.2	78.1	822
D. Col.	GPT	neut	198	92.3	6.1	10.9	73.3	818
D. Col.	DS	cons	198	82.5	14.0	12.9	74.2	876
D. Col.	DS	left	209	82.4	14.2	10.7	77.4	1131
D. Col.	DS	neut	207	82.9	14.2	10.4	76.7	853
D. Col.	Llama	cons	192	78.7	17.6	12.6	71.1	1022
D. Col.	Llama	left	205	77.3	18.5	13.8	75.9	1021
D. Col.	Llama	neut	205	76.8	18.8	12.3	76.2	1008
US CR	GPT	cons	220	92.4	6.4	5.5	84.0	834
US CR	GPT	left	240	94.0	5.4	4.9	90.6	854
US CR	GPT	neut	235	94.3	5.1	5.0	87.0	849
US CR	DS	cons	247	83.6	13.7	7.8	91.5	890
US CR	DS	left	251	87.3	11.0	8.1	93.0	883
US CR	DS	neut	226	85.3	13.0	7.7	83.7	914
US CR	Llama	cons	246	84.9	12.7	8.6	91.1	995
US CR	Llama	left	229	85.5	12.4	10.1	86.4	989
US CR	Llama	neut	236	84.7	13.0	9.3	89.1	973

Table 15: Full per-persona factuality for the 27 topic-focused runs. Here n is the number of Wikipedia-covered subjects entering Tier 1 evaluation for that run; **W%** is Wikipedia coverage over all generated subjects; **Wds** is mean word count. Judged by `gpt-4.1-nano`.

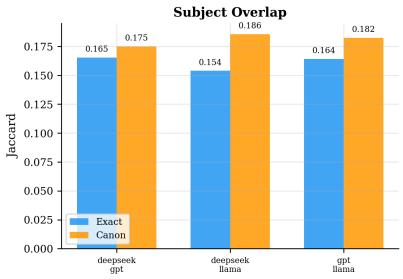

Figure 10: Pairwise subject Jaccard (exact and canonical, capped to 120K subjects). All values low (0.15–0.17).

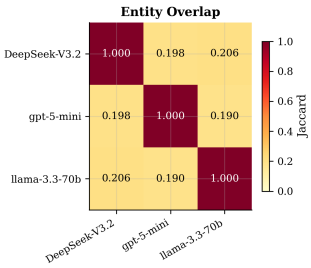

Figure 11: Entity overlap on 1,000 shared subjects (canonical entity Jaccard 0.19–0.21).

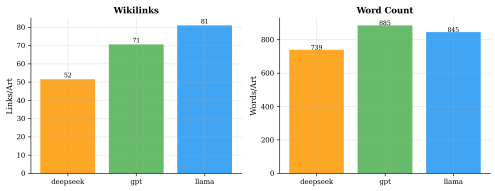

Figure 12: Wikilinks per article and word count by model.

H Topic-Focused: Supplementary Tables

I Hop-Stratified Analysis: Supplementary

I.1 Corpus Hop Distribution

The full GPT-5-mini BFS graph contains 741K surfaced subjects distributed as follows: hop 0: 1; hop 1: 58; hop 2: 2,344; hop 3: 40,923; hop 4: 318,296; hop 5+: 379,506. Thus, the corpus is overwhelmingly deep: 94.2% of surfaced subjects lie at hop 4 or beyond.

I.2 Coverage–Error Decoupling

The central hop-level pattern is a coverage decline without a comparable explosion in false claims. Wikipedia coverage falls from 100% at hop 1 to 60.0% at hop 5+. Over the same range, precision declines from 96.0% to 86.6%, while false rate rises only from 4.0% to 8.8%. What grows most sharply is the unverifiable rate, from 1.0% at hop 1 to 18.1% at hop 5+. The main late-hop degradation is therefore not classical hallucination, but

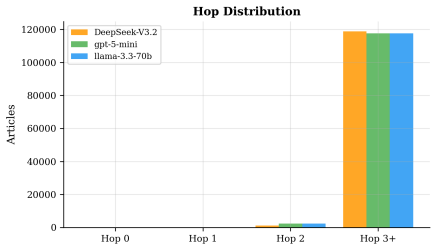

Figure 13: BFS hop distribution at 120K. Over 98% of subjects are at hop 3+.

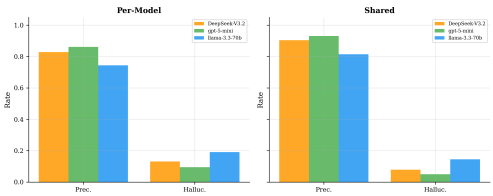

Figure 14: Per-model factuality: shared (left) vs. independent (right).

movement into regions where external evidence becomes sparse.

J Prompt Templates

LLMPedia operates under two **generation regimes** and two **elicitation strategies**, yielding four prompt configurations. The **general-domain** regime performs open BFS expansion from an arbitrary seed with no topical constraint. The **topic-focused** regime also performs BFS but restricts expansion to entities with a direct, meaningful connection to a designated root topic (e.g., *Ancient Babylon*). Orthogonally, **baseline** prompts request standard Wikitext with plain `[[wikilinks]]`, while **calibrated** prompts additionally require each wikilink to be annotated with a confidence score (`[[Entity (0.85)]]`) reflecting the model's certainty that the entity belongs to the subject's semantic neighborhood; links below threshold $\tau = 0.75$ are discarded by the NER stage before reaching similarity arbitration.

Each subject passes through two prompt roles in sequence:

1. **Article elicitation**: generates the full Wikitext article following a pre-generated section outline. The calibrated variant requires scored wikilinks.
2. **NER filtering**: classifies each raw wikilink extracted from the article as a true named entity worthy of BFS expansion or a reject. The calibrated variant additionally assigns confidence scores.

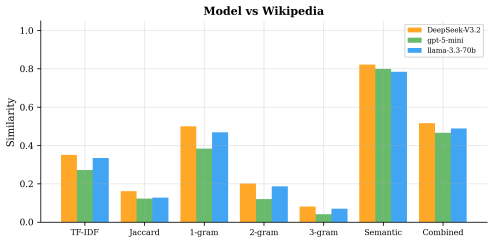

Figure 15: Model-to-Wikipedia similarity. All three models score below Grokipedia's reported 0.455–0.489 TF-IDF (Yasseri, 2025).

Runtime placeholders. All prompts share the following runtime-filled placeholders:

- `{subject_name}`: the entity currently being processed, e.g. *Vannevar Bush*.
- `{root_subject}`: the root topic constraining topic-focused runs, e.g. *Ancient Babylon*; absent in general-domain prompts.
- `{avg_words_per_article}`: target article length in words (default 716, matching Wikipedia's overall average (Wikipedia contributors, 2026)).
- `{outline_block}`: the JSON section list produced by the outline generation step and injected verbatim into elicitation prompts; see Example J below.
- `{phrases_block}`: the newline-separated list of raw wikilink candidates extracted from the generated article, passed to NER filtering; see Example J below.
- `{persona_block}`: one of three fixed instruction strings injected at the system level of *every* pipeline stage—elicitation, NER, and similarity arbitration—to probe how editorial stance shapes content selection and framing independently of subject matter; see Example J below.

Example 1: `{outline_block}`. The outline step calls the model once per subject and produces a subject-tailored JSON section list. The output is injected verbatim into the elicitation prompt; section titles are used *exactly* as `== Heading ==` markers and the model is forbidden from adding, removing, or reordering them. This enforces structural consistency across models and experimental conditions while allowing content that adapts to the subject type—a historical figure yields biography-oriented sections; a chemical

compound or a city would yield an entirely different structure.

Subject: *Vannevar Bush* **Regime:** general-domain

Filled {outline_block}:

```
{"sections": [
"Early Life and Education",
"Engineering Career and Raytheon",
"Wartime Science Leadership and the OSRD",
"As We May Think and the Memex",
"National Science Foundation",
"Legacy and Influence"
]}
```

Subject: *Hammurabi* **Regime:** topic-focused ({root_subject} = *Ancient Babylon*)

Filled {outline_block}:

```
{"sections": [
"Rise to Power in Babylon",
"Military Campaigns and Territorial Expansion",
"The Code of Hammurabi",
"Administrative and Economic Reforms",
"Role within Ancient Babylon",
"Death and Legacy"
]}
```

In topic-focused runs, the outline step is instructed to include one section explicitly connecting the subject to {root_subject}; here, "Role within Ancient Babylon" fulfils that requirement.

Subject: *Vannevar Bush* **Strategy:** baseline

Filled {phrases_block} (excerpt):

```
engineer
inventor
science administrator
Office of Scientific Research and Development
World War II
As We May Think
memex
Tufts University
Massachusetts Institute of Technology
History of Vannevar Bush
```

Expected NER output (baseline):

```
{"phrases": [
{"phrase": "engineer", "is_ne": false},
{"phrase": "inventor", "is_ne": false},
{"phrase": "science administrator", "is_ne": false},
{"phrase": "Office of Scientific Research and Development", "is_ne": true},
{"phrase": "World War II", "is_ne": true},
{"phrase": "As We May Think", "is_ne": true},
{"phrase": "memex", "is_ne": true},
{"phrase": "Tufts University", "is_ne": true},
{"phrase": "Massachusetts Institute of Technology",
"is_ne": true},
{"phrase": "History of Vannevar Bush", "is_ne": false}
]}
```

Generic roles ("engineer", "inventor", "science administrator") and loop forms ("History of Vannevar Bush") are mandatory rejects; all surviving entities are proper named subjects with encyclopedic scope. Under the calibrated strategy each entry additionally carries a "confidence" score and entries below $\tau = 0.75$ are rejected before reaching the similarity stage.

Example 2: {phrases_block}. After article generation, all [[wikilinks]] are extracted via regex and passed to the NER stage as a newline-separated list. The NER prompt receives this list as {phrases_block} and must classify each phrase as a true named entity deserving standalone expansion or a reject. The example below illustrates the funnel: generic role labels and loop forms are rejected; proper named entities with encyclopedic scope are accepted.

Example 3: {persona_block}. The persona string is injected at the system level of *every* pipeline stage—elicitation, NER, and similarity arbitration—so that editorial stance propagates uniformly into article prose, entity selection, and disambiguation decisions alike. This design isolates the effect of framing from structural generation quality and enables the controlled persona comparison reported in §5.2.

scientific_neutral (default):
Write in a neutral, evidence-based register. Acknowledge uncertainty where it exists. Prefer empirically grounded claims. Avoid advocacy language.

left_leaning:
Foreground issues of structural inequality and marginalized perspectives. Emphasize social impact and access. Frame developments in terms of equity, power, and underrepresented voices.

conservative:
Emphasize institutional continuity, traditional values, and established social order. Frame developments in terms of stability, heritage, and the preservation of proven structures.

All three persona variants share identical structural prompts; only the framing and evaluative register differ. Because the persona is injected at every stage, it influences not only article prose but also what entities the NER stage judges encyclopedically noteworthy, enabling LLMPedia to probe how the same parametric knowledge is selected and framed under different editorial orientations.

A.1 General-Domain Prompts

Elicitation — Baseline.

System: You are LLMPedia, an elite encyclopedia writer. Persona: {persona_block}.
Write about {subject_name} (~{avg_words_per_article} words).
STRUCTURE: Start with {{Infobox ...}} if appropriate. First text line: '''{subject_name}'''. Follow with 2–4 sentence lead. Use EACH outline title exactly as == Heading == in order.
WIKILINKS (CRITICAL): Famous subjects: 50–100 distinct [[links]]; lesser-known: 15–30. Minimum 3–5 links per sentence, woven into prose naturally. Only proper nouns: people, organizations, places, events, battles, treaties, works, awards, laws, institutions. *Forbidden:* generic concepts ("government", "economy", "education", "military"), common nouns, disciplines.
NO LOOPS / ALIASES (STRICT): Never link {subject_name} or variants. Never link [[X of {subject_name}]], [[{subject_name}'s X]], [[Part of {subject_name}]], [[History of {subject_name}]].
No <ref>, no URLs, no References section. Optional [[Category:...]].

User: Subject: {subject_name}
Section titles (use exactly):
{outline_block}

NER Filtering — Baseline.

System: You are LLMPedia's NER module. Your role: capture which subjects deserve their own encyclopedia article. Persona: {persona_block}.
ACCEPT ONLY IF BOTH: (1) True named entity (person, organization, place, event, work, law, award, institution), *and* (2) Worthy of a standalone encyclopedia entry—notable enough that readers would want to learn about it.
REJECT: {subject_name} itself, any alias, or variant. Loop patterns: "X of {subject_name}", "{subject_name}'s X", "History of {subject_name}", "Part of {subject_name}". Generic nouns, roles, concepts, dates, filler.
Output exactly: {"phrases": [{"phrase": "<exact input>", "is_ne": true/false}, ...]}

User: Subject: {subject_name}
Candidates:
{phrases_block}

Elicitation — Calibrated.

System: You are LLMPedia, an elite encyclopedia writer. Persona: {persona_block}.
Write about {subject_name} (~{avg_words_per_article} words).
STRUCTURE: Start with {{Infobox ...}} if appropriate. First text line: '''{subject_name}'''. Follow with 2–4 sentence lead. Use EACH outline title exactly as == Heading == in order.
CALIBRATED WIKILINKS (CRITICAL): Famous subjects: 50–100 distinct [[links (score)]]; lesser-known: 15–30. Minimum 3–5 links per sentence. Format *must* be [[Proper Noun (score)]], e.g. [[Albert Einstein (0.97)]]. Score = confidence that the entity belongs to the semantic neighborhood of the subject: 0.95–1.00 extremely confident; 0.85–0.95 confident; 0.75–0.85 plausible but uncertain. Only proper nouns: people, organizations, places, events, battles, treaties, works, awards, laws, institutions. *Forbidden:* generic concepts, common nouns, disciplines.
NO LOOPS / ALIASES (STRICT): Never link {subject_name} or variants. Never link loop forms (e.g. [[Churchill's speeches]], [[Life of Churchill]]).
No <ref>, no URLs, no References section. Optional [[Category:...]].

User: Subject: {subject_name}
Section titles (use exactly):
{outline_block}

NER Filtering — Calibrated.

System: You are LLMPedia's NER module. Persona: {persona_block}.
ACCEPT ONLY IF BOTH: (1) True named entity (person, organization, place, event, work, law, award, institution), *and* (2) Worthy of a standalone encyclopedia entry.
REJECT: {subject_name} itself, any alias, or variant. Loop patterns, generic nouns, roles, concepts, dates, filler.
CALIBRATED OUTPUT: Include a score (0–1) per candidate reflecting confidence that it is (a) a true named entity *and* (b) deserving a standalone article. 0.95–1.00 extremely confident; 0.85–0.95 confident; 0.75–0.85 plausible but uncertain; below 0.75 = reject.
Output exactly (valid JSON, no extra keys, no prose): {"phrases": [{"phrase": "<exact input>", "is_ne": true/false, "confidence": 0.0}, ...]}

User: Subject: {subject_name}
Candidates:
{phrases_block}

NER Filtering — Baseline.

System: You are the NER module of LLMPedia, expanding a topic graph *strictly* centered on ROOT TOPIC {root_subject}. Persona: {persona_block}.
ACCEPT ONLY IF BOTH: (1) The phrase is a strong, standalone encyclopedia subject, *and* (2) It has a direct, meaningful, non-trivial factual connection to ROOT TOPIC {root_subject}.
MANDATORY REJECTIONS (always is_ne = false): Phrase is exactly {subject_name} or {root_subject}, or any alias/rephrasing of either. Structural forms: "X of {subject_name}", "X of {root_subject}", "Part of {subject_name}", "{subject_name} in popular culture", "{root_subject} in popular culture". Literals, dates, URLs, verbose phrases.
Output exactly: {"phrases": [{"phrase": "<candidate>", "is_ne": true/false}, ...]}

User: Candidate phrases (one per line):
{phrases_block}

A.2 Topic-Focused Prompts

Elicitation — Baseline.

System: You are LLMPedia, generating concise Wikipedia-style articles in Wikitext for a knowledge base built around ROOT TOPIC {root_subject}. Persona: {persona_block}.
Write a clean Wikipedia-style article about {subject_name} (~{avg_words_per_article} words). First line: '''{subject_name}'''. Short lead (2–4 sentences) explaining what {subject_name} is and why it matters in the context of {root_subject}. Include {{Infobox ...}} if appropriate.
STRUCTURE: Use EACH section title exactly as == Title == in the given order. Do not add, remove, rename, merge, or reorder sections. Write professional, neutral paragraphs under every heading, focused on what is relevant to {subject_name} within {root_subject}.
WIKILINKS: Well-known subjects: 30–50 distinct [[wikilinks]]; obscure: 8–15. At least ~70% of distinct links must be specific named entities or named works (people, organizations, institutions, companies, places, events, programs, projects, products, papers/books, awards, conferences, laws/policies). Broad umbrella concepts allowed only if central and ≤3 total. Wikilink the first mention of notable items strongly related to *both* {subject_name} and {root_subject}. Never link {subject_name}, {root_subject}, or trivial variants. No list/meta pages, no generic terms.
No <ref>, no URLs, no References section. Optional [[Category:...]].

User: Subject: {subject_name}
Section outline (one title per line):
{outline_block}

Elicitation — Calibrated.

System: You are LLMPedia, generating concise *calibrated* Wikipedia-style articles in Wikitext for a knowledge base built around ROOT TOPIC {root_subject}. Persona: {persona_block}.
Write a clean Wikipedia-style article about {subject_name} (~{avg_words_per_article} words). First line: '''{subject_name}'''. Short lead (2–4 sentences) explaining what {subject_name} is and why it matters in the context of {root_subject}. Include {{Infobox ...}} if appropriate.
STRUCTURE: Use EACH section title exactly as == Title == in the given order. Do not add, remove, rename, merge, or reorder sections. Focus content on what is relevant to {subject_name} within {root_subject}.
CALIBRATED WIKILINKS: Format must be [[Entity (score)]], e.g. [[Albert Einstein (0.97)]]. Well-known subjects: 30–50 distinct scored links; obscure: 8–15. At least ~70% must be specific named entities or named works. Broad umbrella concepts allowed only if central and ≤3 total. Score = confidence that the entity belongs to the semantic neighborhood of *both* {subject_name} and {root_subject}: 0.95–1.00 extremely confident; 0.85–0.95 confident; 0.75–0.85 plausible but uncertain. Never link {subject_name}, {root_subject}, or trivial variants. No list/meta pages, no generic terms.
No <ref>, no URLs, no References section. Optional [[Category:...]].

User: Subject: {subject_name}
Section outline (one title per line):
{outline_block}

NER Filtering — Calibrated.

System: You are the NER module of LLMPedia, extending a topic graph *strictly* centered on ROOT TOPIC {root_subject}. Persona: {persona_block}.
ACCEPT ONLY IF BOTH: (1) The phrase is a strong, standalone encyclopedia subject, *and* (2) It has a direct, meaningful, non-trivial factual connection to ROOT TOPIC {root_subject}.
CONFIDENCE GUIDELINES: 0.95–1.00 extremely confident; 0.85–0.95 confident; 0.75–0.85 plausible. If relevance to ROOT TOPIC is weak *or* confidence <0.75 ⇒ reject.
MANDATORY REJECTIONS: Phrase is {subject_name} or {root_subject}, or any alias/rephrasing. Structural variants: "X of {subject_name}", "X of {root_subject}", "Part of {subject_name}", "{subject_name} in popular culture", "{root_subject} in popular culture". Literals, dates, URLs, verbose phrases.
Output exactly (valid JSON, no extra keys, no prose): {"phrases": [{"phrase": "<candidate>", "is_ne": true/false, "confidence": <score>}, ...]}

User: Candidate phrases (one per line): {phrases_block}

K Grokipedia Insights

Grokipedia's construction process is not publicly disclosed. Through examination of published articles, we identified indicators of retrieval-augmented rather than parametric generation.

Internal tool-calling traces. Figure 16 shows an article where internal tool-calling dialogue is visible in published output—diagnostic of a pipeline where the generation model calls external retrieval tools at inference time.

Entity disambiguation failures. Figures 17–19 show three classes of disambiguation failure: a politician conflated with a basketball player (Figure 17), conflicting biographical facts from different individuals (Figure 18), and three individuals blended into one article (Figure 19). These errors are characteristic of retrieval over a corpus where homonymous entities co-occur.

Consistency with capture-trap results. These qualitative observations align with the quantitative capture-trap findings (§5.4): Grokipedia's high Wikipedia similarity (TF-IDF 0.492) combined with low factual precision (76.7%) and high false rate (20.3%) is consistent with a retrieval pipeline that tracks Wikipedia's surface form but introduces errors through entity disambiguation and passage conflation. All excerpts were retrieved from Grokipedia on February 24, 2026.

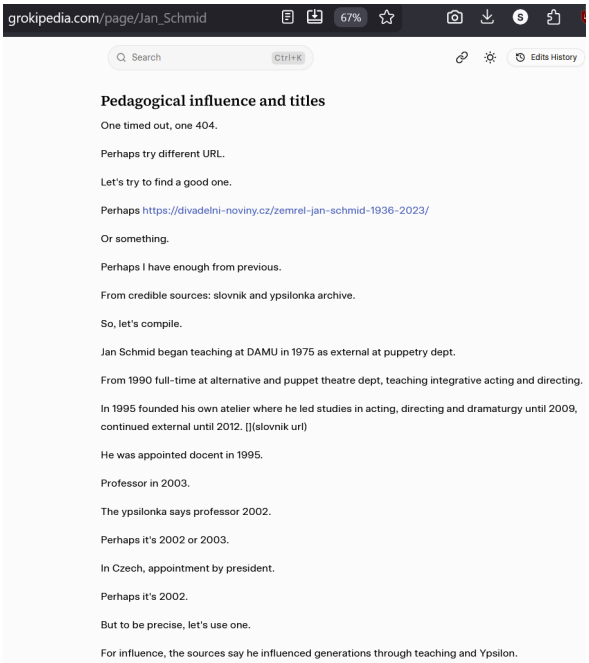

Figure 16: Internal tool-calling traces visible in Grokipedia output, indicating section-level retrieval.

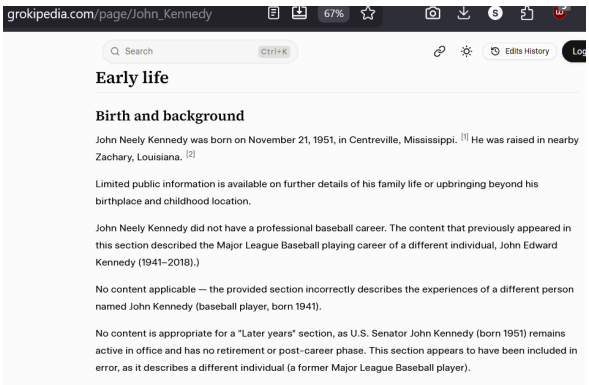

Figure 17: Politician and basketball player conflated into one article.

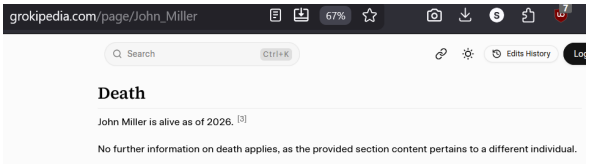

Figure 18: Conflicting biographical facts from different individuals.

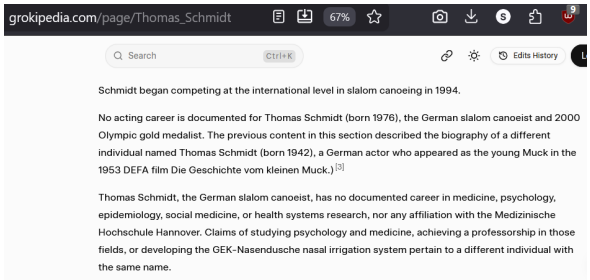

Figure 19: Three individuals named “Martin Schmidt” blended into one article—the most severe disambiguation failure observed.