

MUSART: A Scalable Multi-Domain Benchmark for Measuring Factual Recall in Large Language Models

Rajaa El Hamdani,¹ Thomas Bonald,¹ Fragkiskos Malliaros^{2,3}

¹Institut Polytechnique de Paris

²Université Paris-Saclay, CentraleSupélec, ³Inria

rajaa.elhamdani@telecom-paris.fr, thomas.bonald@ip-paris.fr

fragkiskos.malliaros@centralesupelec.fr

Abstract

Reliable evaluation of factual recall in large language models (LLMs) requires ground truth that is broad, clean, and complete. Yet existing benchmarks often rely on narrow domains or incomplete knowledge bases that penalise models for generating correct facts absent from the reference. We introduce **MUSART**, a benchmark of 21,775 subject–relation questions over 11 domains, created via a schema-free extraction pipeline that can be cheaply regenerated from public dumps. To overcome knowledge-base incompleteness, we contribute a source-grounded augmentation pipeline that mines verifiable facts directly from article text. Evaluating twenty-one open-weight LLMs against a dual-reference pair (original vs. augmented) yields three findings: (i) completeness correction systematically raises precision and lowers recall across all models; (ii) factual recall varies widely across subject domains, demonstrating that single-domain evaluation confounds model capability with subject matter; and (iii) the precision gap between the two references separates instances where models lack knowledge from instances where the knowledge base lacks facts, proving that completeness-blind evaluation penalises both identically. All code, data, and evaluation scripts are publicly released at <https://github.com/Rajjaa/MUSART>.

1 Introduction

Evaluating large language models as knowledge bases (LM-as-KB) is fundamentally confounded by the references used to score them. Existing benchmarks suffer from two structural blind spots: narrow domain coverage, which locks in arbitrary difficulty baselines, and knowledge-base incompleteness, which penalises models for generating correct facts absent from the ground truth. For instance, when an LLM correctly names an ensemble cast member of a television series whom Wikipedia documents but Wikidata omits, standard evaluation

records a false positive—marking the model wrong for being right.

We address both challenges with **MUSART** (Multi-domain SAlient Relation Triples). Specifically, our contributions are threefold:

1. **A multi-domain factual benchmark:** We release 21,775 subject–relation questions across 11 domains, created via a schema-free extraction pipeline that can be cheaply regenerated from public dumps.
2. **A dual-reference evaluation paradigm:** To mitigate reference incompleteness, we contribute a source-grounded augmentation pipeline that mines verifiable facts directly from article text. We propose evaluating models against a conservative–optimistic pair of references to expose missing facts.
3. **A large-scale empirical study:** We evaluate twenty-one open-weight LLMs against this dual-reference setup, isolating factual recall from evaluation artefacts.

Our evaluation yields three principal findings:

1. **Completeness correction is systematic:** Moving from the original to the augmented ground truth raises precision and depresses recall in every single run, showing that reference incompleteness is a major, predictable confounder in factuality scoring (Section 5).
2. **Domain variance is large and shared:** Factual recall varies widely across subject domains (from 50.1% down to 10.4% mean precision), and models show strong agreement on which subjects are hard (mean pairwise Spearman $\rho = 0.68$). This suggests the variation reflects subject matter and the shared inductive biases of decoder-only Transformers

rather than individual model behavior, demonstrating that single-domain evaluation cannot serve as a general study of parametric knowledge (Section 5.3).

3. **Correction separates model failure from KB failure:** The precision gap between Wikidata and article text varies dramatically across domains (from 14.9 to 0.6 points). This variance isolates where models lack knowledge versus where the knowledge base lacks facts, proving that completeness-blind evaluation penalises both identically (Section 5.3).

All code, data, question templates, and evaluation scripts are publicly available at <https://github.com/Rajjja/MUSART>.

2 Related Work

Factual Recall and Knowledge Bases. The probing paradigm established by LAMA (Petroni et al., 2019) has scaled from single-token cloze tests to full generative benchmarks (Jiang et al., 2020; Elazar et al., 2021; Hu et al., 2024). Triple-aligned resources like T-REx (Elsahar et al., 2018) and KAMEL (Kalo and Fichtel, 2022) broadened relation coverage, while benchmarks like PopQA (Mallen et al., 2023), Head-to-Tail (Sun et al., 2024), and SimpleQA (Wei et al., 2024) select facts based on entity popularity. Popularity-based selection risks under-sampling domain-specific knowledge; MUSART instead seeds from thematic Wikipedia categories and ranks relations by graph salience to ensure broad multi-domain coverage.

Evaluation under Imperfect Ground Truth. Knowledge bases follow the open-world assumption, meaning absent triples generate false negatives during evaluation (Galárraga et al., 2013). Existing approaches handle this through statistical correction of label noise (Jain et al., 2016; Ramola et al., 2019) or calibrated-judge frameworks that estimate error rates (Feng et al., 2026). MUSART takes a different approach: rather than producing a single corrected point estimate, it evaluates models against a conservative–optimistic pair of references. Furthermore, while recent benchmarks increasingly rely on LLMs as judges (Min et al., 2023; Zheng et al., 2023) or embedding metrics (Zhang et al., 2020), MUSART holds the scorer fixed and deterministic via dense alias matching. This ensures that any score movement between references is attributable to knowledge-base completeness rather

than grader instability. Finally, because MUSART is strictly single-hop, it cleanly isolates factual recall from the complex effects of test-time reasoning (Wei et al., 2022; Snell et al., 2024; Chen et al., 2025; Yao et al., 2025).

3 The MUSART Benchmark

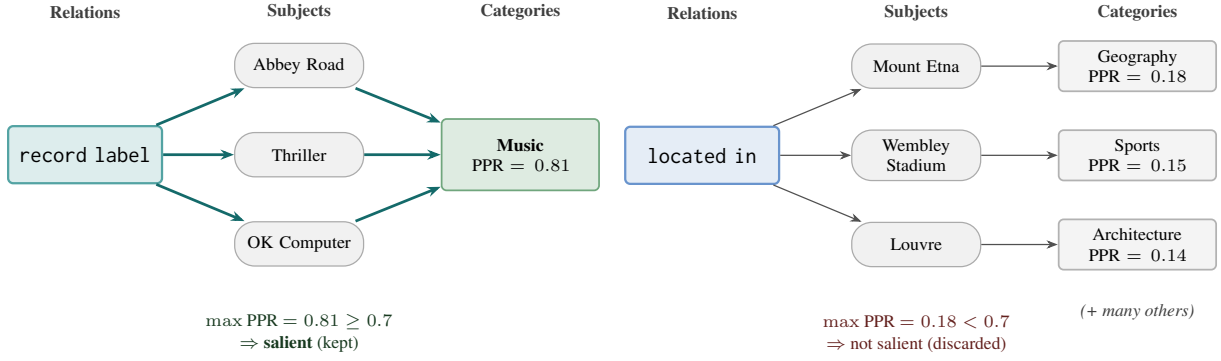
3.1 Triple Acquisition and Filtering

The initial phase of benchmark construction identifies high-quality source articles, links them to Wikidata, retrieves associated structured triples, and rigorously filters them before computing domain salience (see Appendix A for full technical details):

- **Source Article Selection:** We extract 38,279 English Wikipedia Good Articles (GAs) across 65 second-level domain categories (e.g., “Science and technology” → “Biology and medicine”).
- **Wikidata Linking:** Using Wikimedia dump mappings, we align GA titles to Wikidata QIDs, excluding 11,353 unmapped articles to retain 26,926 linked entities, of which 26,210 yield statement triples in Wikidata.
- **Triple Acquisition:** We extract 302,179 truthy statement triples (s, r, o) across 1,001 relations from Wikidata dumps associated with these 26,210 subjects.
- **Data Filtering & Salience:** We remove 24 generic or administrative relations (Table 6 in Appendix D) along with ID properties, and apply a frequency threshold (≥ 50 occurrences) to prune 689 sparse relations. Subsequent Personalized PageRank salience selection (Section 3.2) is followed by manual domain selection and cardinality filtering to isolate the final benchmark of 44,722 triples across 8,407 subject–article pairs and 37 salient relations over 11 domains, grouped into 21,775 set-valued questions (Section 5).

3.2 Saliency Scoring

Salient Relation Identification. The core of our methodology is to automatically identify relations that are particularly salient or characteristic for each domain. The goal is to filter out generic relations that appear uniformly across many categories—such as “designed by” or “author”—and to retain relations that are specific to the semantic field of a category. For instance, an ideal system



(a) Domain-specific relation: PPR mass *concentrates* on one category.

(b) Generic relation: PPR mass is *spread* across many categories.

Figure 1: The graph-based salience model behind MUSART, contrasting the two regimes it must separate. A heterogeneous graph links each relation to the subjects in which it occurs and each subject to its Wikipedia category; a Personalized PageRank walk is initiated from each relation node. (a) For a *domain-specific* relation such as *record label*, the subjects overwhelmingly belong to one category, so the walk concentrates probability mass there (Music, 0.81) and the relation is retained as salient. (b) For a *generic* relation such as *located in*, the subjects are scattered across many domains, so the mass is diluted and no single category exceeds the global threshold (0.7); the relation is discarded. This is precisely the distinction that purely frequency-based scoring fails to make.

should identify that the relation “record label” is highly salient for music-related categories, but not for geography. Table 4 in Appendix A presents a set of manually curated examples that illustrate the desired quality and domain-specificity of the relations we aim to extract automatically.

Manually creating such lists for all domains is not scalable. Therefore, we developed an automated pipeline to approximate this level of quality. We first explored a simple frequency-based statistical approach, but found it had significant limitations, which led us to develop a more robust graph-based method using Personalized PageRank (PPR) for the final dataset construction.

Baseline: Frequency-based Saliency. We initially scored each relation r in category c by its in-domain frequency ratio: $\text{Saliency}(r, c) = \frac{\text{count}(r, c)}{\text{count}(r)}$. While an empirical threshold of 0.2 balanced specificity and coverage for certain categories, score distributions varied wildly across domains. Consequently, a single global threshold left several domains with zero salient relations, proving unworkable without non-scalable, per-category manual tuning.

Final Method: Personalized PageRank-based Saliency. To overcome the limitations of the frequency-based baseline, we employ a graph-based method using Personalized PageRank (PPR). The intuition is that relations disproportionately used within a specific set of categories will primar-

ily link to subjects belonging to those categories. Therefore, by constructing a graph representing these connections (Relation $\rightarrow$ Subject $\rightarrow$ Category), a random walk starting from such a relation is expected to frequently visit the nodes representing its most relevant categories.

We construct a heterogeneous directed graph $G = (V, E)$ from the filtered Wikidata triples and their associated Wikipedia categories, where nodes represent subjects, relations, and second-level Wikipedia categories, and edges link subjects to their categories and relations to their subjects (see Appendix A for formal node-set and edge-set definitions). We run PPR from each unique relation r to compute importance scores across all category nodes. Sorting relations by this score yields a ranked list for each category, where an empirical threshold of 0.7 effectively separates highly relevant domain relations from peripheral ones. Figure 1 illustrates the intuition behind this graph-based salience model. Table 1 summarises the resulting dataset statistics before and after salience selection.

Qualitatively, PPR recovers domain-specific relations (e.g., hymenium attachment for *Biology and medicine*; MPA film rating for *Film*). Nine categories yield no salient relations—five of them sports-related—because their GA representation relies on shared generic properties, validating the method’s ability to distinguish topical relevance from true domain salience. From the 41 Good-

	Initial	Final
Articles	38,279	8,407
Categories	65	11
Triples	302,179	44,722
Subjects	26,210	8,407
Relations	1,001	37
Questions	—	21,775

Table 1: MUSART statistics. *Initial* = after triple acquisition, before relation filtering and PPR salience selection; *Final* = the released benchmark, after salience selection, manual domain selection and cardinality filtering. Triples sharing a subject–relation pair are grouped into one set-valued question, giving 2.05 gold objects per question.

Articles categories retained by the saliency step, we manually selected a subset of domains for evaluation to limit the scope of the paper, as the full dataset would be too large to evaluate and study manually, yielding 11 domains after relation-frequency and cardinality filtering.

4 Ground-Truth Augmentation

4.1 Motivation: Knowledge-Base Incompleteness

Wikidata is not exhaustive, especially for high-cardinality relations (e.g. a film’s full cast). An LLM may produce a correct fact that is absent from the ground truth and be wrongly penalised, systematically underestimating precision. We therefore enrich the Wikidata ground truth with facts mined from the GA article text.

4.2 Augmentation Pipeline

To enrich Wikidata without sacrificing transparency, we perform extractive question answering on the Good Article texts using `langextract` with Gemini 3.1 Flash Lite. We filter extractions to retain only verbatim, character-grounded quotes (Tier 1 candidates), mapping hyperlinked spans to QIDs via Wikimedia dumps and retaining unlinked surface forms after string deduplication (see Appendix B and Figure 5 for full technical specifications). Processing 21,662 article–relation pairs yields 69,208 grounded extractions (61.4% QID-linked), raising the mean gold objects per question from 2.05 to 4.13.

Reading the Two Ground Truths. This pipeline yields two complementary reference regimes: (1)

the **original ground truth** (Wikidata-only), a *conservative* reference that penalises correct facts missing from Wikidata; and (2) the **augmented ground truth**, an *optimistic* reference that credits article-supported answers while inheriting unlinked noise and deduplication limits (detailed in Appendix B, Table 5). Together, they measure factual sensitivity to reference completeness rather than point estimates of precision.

5 Evaluating Large Language Models

5.1 Experimental Setup

The evaluation set comprises **21,775 questions** spanning eleven domains (3,373 single-valued, 18,402 multi-valued; 2.05 gold objects/question on average), converted from structured triples via relation-specific templates following the LM-as-KB paradigm (Petroni et al., 2019). Models are evaluated under five-shot prompting without access to source texts (testing parametric knowledge only) and scored using set-based *alias matching* (AM), where an object is recalled if and only if any of its surface forms is generated as an exact string match (substring matches are rejected). All metrics (precision, recall, and F1) take values in $[0, 1]$ and are reported as percentages (%), micro-averaged against both reference regimes; full template generation, serving, and scoring mechanics are detailed in Appendix G.

Models and Evaluation Scope. While the MUSART benchmark and its dual-reference evaluation paradigm are model-agnostic and apply equally to closed APIs, our empirical study focuses on twenty-one open-weight models, spanning 0.8B to 70B parameters across four families (Qwen3.5, Gemma-2/Gemma-3/Gemma-4, Llama-3, and Mistral/Minstral). Using open-weight models ensures reproducibility and execution without proprietary API drift or unannounced backend modifications. The slate systematically varies three usually confounded factors: *scale* within a family and generation (e.g., Llama-3 from 1B to 70B), *generation* at matched size (Gemma-2, 3, and 4 across 2B–9B scales), and architectural *family* at matched scale in the 7–12B band (see Appendix G for full slate details). While main results use direct-answer configurations supported across all twenty-one models, thinking modes for the Qwen3.5 and Gemma-4 series are evaluated separately in Appendix F.

5.2 Completeness Correction: Precision Up, Recall Down

Evaluating against the augmented (vs. original) ground truth raises precision but lowers recall (Table 3). For Llama-3.3 70B, precision rises from 39.71% on the original ground truth to 46.47% on the augmented one, while recall falls from 52.47% to 30.53%. The direction is uniform: precision rises and recall falls in every row of Table 3. Figure 6 shows the resulting spread for every model; its width is how far that model’s measured precision moves when the reference changes. Precision is the right cross-ground-truth comparator; recall and F1 fall mechanically because the augmented reference roughly doubles the gold objects to enumerate while models emit roughly the same number of answers. Because the augmented ground truth is un-audited and admits extraction and linking noise (Section 4), this rise is *consistent with*—but does not by itself establish—genuine underestimation of model precision by the original benchmark: part of the gain may come from answers that match noisy unlinked spans. The recall drop, by contrast, reflects the larger object count of the augmented set (2.05 to 4.13 gold objects per question; Section 4), including unlinked objects that may duplicate existing ones, and exposes a long-tail gap between what models produce and what the source text supports.

Where the reference gap is widest. Splitting the evaluation set by answer cardinality localises both effects (Table 2). Single-valued questions (3,373) are easier than multi-valued ones (18,402) under either reference—averaged over the slate, F1 is 5.99 points higher on the original ground truth and 7.59 points higher on the augmented one—but the interesting quantity is how augmentation moves each subset. Precision gains are more than three times larger on multi-valued questions (+3.04 points on average, against +0.90 for single-valued), and the recall penalty is also larger there (−12.81 against −8.97). Both movements point the same way: high-cardinality relations are where Wikidata is most incomplete, so they are simultaneously where the original benchmark most under-credits correct answers and where augmentation adds the most gold objects to enumerate. The reference gap is therefore not uniform; it is widest exactly where the knowledge base is thinnest, which is also where a completeness-blind evaluation is most misleading.

		Single val.	Multiple val.
Count		3,373	18,402
Precision	original	33.40	28.63
	augmented	34.30	31.67
	Δ	+0.90	+3.04
Recall	original	35.21	30.09
	augmented	26.24	17.27
	Δ	−8.97	−12.81

Table 2: The reference gap by answer cardinality: alias-match precision and recall (%) against each ground truth, averaged over the twenty-one models. Both effects of completeness correction are larger on multi-valued questions—the precision gain is more than three times larger, and the recall penalty is also steeper—reflecting where a knowledge base is most likely to be missing objects.

5.3 Factual Recall Varies Widely Across Subject Domains

MUSART’s eleven evaluated domains reveal that measured factual recall is highly sensitive to the subject matter being queried. On the original reference, mean precision varies from 50.1% for *Physics and astronomy* down to 10.4% for *Transport*, and a similarly wide dispersion holds under the augmented one. We analyse precision rather than F1 throughout this section: the augmented reference roughly doubles the gold objects per question, so recall and F1 fall mechanically, whereas a precision gain means an answer the model actually produced is now credited (Section 5.2). This wide cross-domain variance is an empirical observation and a consequence of how pre-training data density and knowledge-base coverage differ by topic. Because baseline difficulty varies so dramatically from subject to subject, evaluating factuality on a single domain locks in an arbitrary difficulty level that cannot serve as a general measure of a model’s factual recall.

This domain profile is a shared property across architectures, not a product of individual model behavior. Figure 2 shows every model–domain cell under both ground truths: the columns are visibly banded, and the domain rankings of any two models agree at mean pairwise Spearman $\rho = 0.67$ on the original ground truth (0.62 on the augmented). Models largely find the same subjects easy and the same subjects hard, which points to the composition of pre-training data and of Wikidata coverage rather than to model-specific gaps. The exceptions are legible in the same figure and are informative:

Model	Original GT			Augmented GT		
	P	R	F1	P	R	F1
Qwen3.5 0.8B	16.90	8.75	11.53	18.68	4.81	7.65
Qwen3.5 2B	21.42	15.16	17.76	23.63	8.32	12.30
Qwen3.5 4B	22.54	30.91	26.07	25.06	17.09	20.32
Qwen3.5 9B	26.53	35.42	30.34	29.33	19.48	23.41
Gemma-3 1B	15.13	12.19	13.50	16.57	6.64	9.48
Gemma-3 4B	24.15	29.64	26.61	26.79	16.35	20.30
Gemma-3 12B	34.62	36.11	35.35	38.63	20.04	26.39
Gemma-3 27B	31.50	43.83	36.65	36.49	25.25	29.85
Gemma-4 E2B	25.83	18.04	21.24	27.66	9.61	14.26
Gemma-4 E4B	30.75	24.12	27.04	34.01	13.26	19.08
Gemma-4 12B	30.12	34.70	32.25	33.74	19.33	24.58
Gemma-4 31B	42.46	40.01	41.20	47.63	22.32	30.40
Gemma-2 2B	29.85	27.84	28.81	33.31	15.45	21.11
Gemma-2 9B	35.01	39.05	36.92	39.34	21.82	28.07
Llama-3.2 1B	15.04	13.51	14.24	16.22	7.25	10.02
Llama-3.2 3B	28.11	29.81	28.93	31.00	16.35	21.41
Llama-3.1 8B	34.19	38.75	36.32	38.61	21.76	27.84
Llama-3.3 70B	39.71	52.47	45.21	46.47	30.53	36.85
Ministral-8B	25.46	29.68	27.41	28.18	16.34	20.68
Mistral-7B v0.3	24.00	35.65	28.69	27.18	20.07	23.09
Mistral-Nemo	32.27	32.86	32.56	35.97	18.21	24.18

Table 3: Main results (alias-match micro-averages, %; few-shot, direct-answer). Note: F1 is comparable only *within* each ground truth; compare precision across ground truths.

Ministral-8B ranks 11th of 21 overall but first on *Physics and astronomy*, where Mistral-Nemo is second, while Llama-3.3 70B—first overall—places ninth in that domain, and Gemma-3 27B falls from fourth overall to eleventh. On a small domain (113 questions) this is an observation rather than a result, but it is the kind of pre-training signature that a single-domain benchmark cannot expose at all.

Where the knowledge base is incomplete. Comparing references per domain estimates Wikidata incompleteness as models experience it (Figure 7 in Appendix C). This gap varies widely across measured domains (Figure 3), ranging from 14.9 points in *Architecture* down to 0.6 in *Motorsport* (where functional relations like pole position admit a single correct answer) and 0.5 in *Biology and medicine*—a thirty-fold spread across the ten domains where augmentation ran. Consequently, knowledge-base incompleteness is highly domain-dependent, and a single global completeness correction would be inaccurate in both directions.

What predicts a wide gap is less clear than we expected. Across the ten domains where augmentation ran, the gap correlates most with the proportion of multi-valued questions ($r = 0.57$), then with the

share of extracted mentions that hyperlink matching resolves to QIDs ($r = 0.47$), then with the factor by which augmentation multiplies the gold-object count ($r = 0.35$). With ten points none of these is decisive, and we report them as indicative rather than as an attribution.

Precision gain and recall loss are almost unrelated across domains ($r = 0.11$), and that dissociation is informative: they measure different incompletenesses. *Architecture* and *Biology and medicine* lose comparable recall (-13.4 and -15.9) but gain wildly different precision ($+14.9$ against $+0.5$). In *Architecture* the article text supplies objects that models already produce and Wikidata simply lacked; in *Biology and medicine* augmentation adds many objects, but ones the models do not know. The extreme case is *Battles, exercises, and conflicts*, where augmentation raises the gold set from 2.5 to 12.0 objects per question—a $4.8\times$ multiplication that produces both a wide gap and the largest recall penalty on the benchmark. The first pattern is a deficiency of the knowledge base, the second a deficiency of the models, and a completeness-blind evaluation reports both as the same low score.

Rank stability. Despite these shifts in level, the domain *ordering* is robust to the choice of ground truth. Computing the correlation separately for each model rather than aggregating (Figure 2, right), per-model Spearman ρ between the original and augmented domain rankings is positive and high throughout: 0.98 for Qwen3.5 0.8B at the top and 0.76 for Ministral-8B at the bottom, with a mean of 0.89. No model falls below 0.75. Since one of the two references is known to carry extraction and linking noise, agreement this high indicates that the domain profile reflects genuine variation in stored knowledge rather than an artefact of missing Wikidata entries.

5.4 Ranking Instability

Changing the reference also changes the leaderboard, though modestly. Between the original and augmented ground truths, seven of the twenty-one models change rank, and the largest move is two places (Gemma-4 E4B). The top of the table is stable—Llama-3.3 70B leads under either reference—and so is the bottom; the movement is mid-table, which is where leaderboard comparisons are usually contested and where the margins are thinnest. We note the effect but do not lean on it: a two-place shift on this slate corresponds to

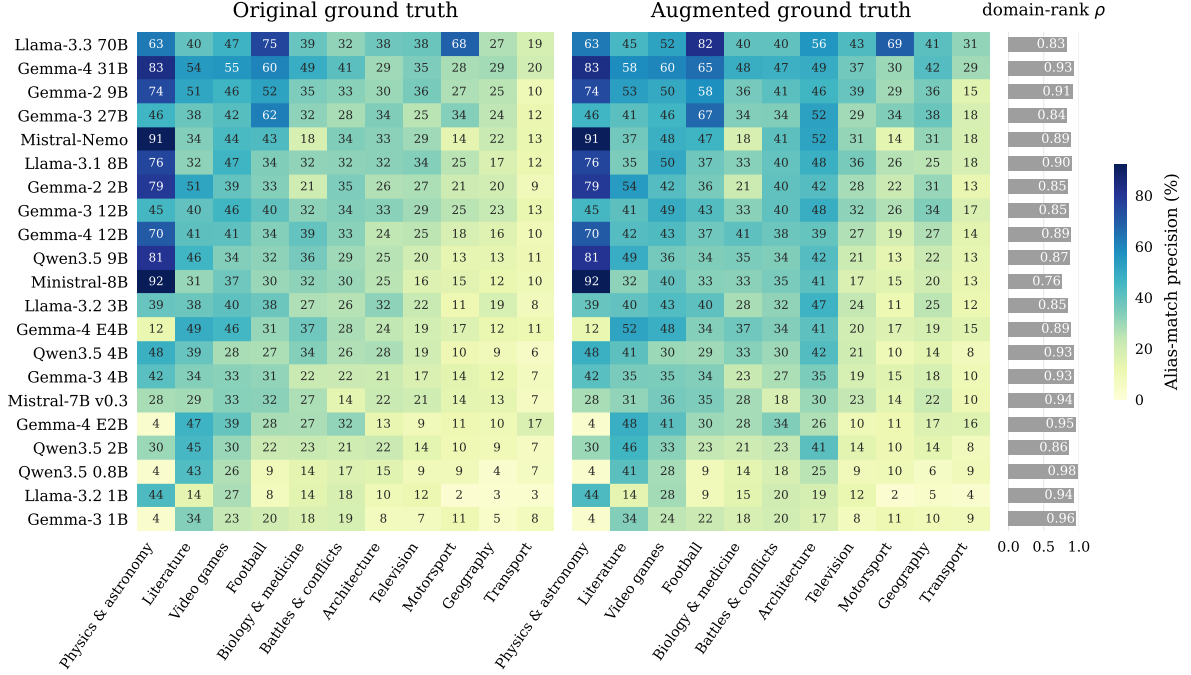


Figure 2: Alias-match precision for every model–domain pair under both references, with models and domains ordered by their original-reference means. Precision is shown because it is the quantity comparable across references: an extracted object the model also produced is very likely a genuine fact the knowledge base was missing, whereas recall falls by construction as the gold set grows. Right: each model’s Spearman correlation between its original and augmented domain rankings. Columns are visibly banded across subjects—reflecting how baseline difficulty varies by domain—and the per-model ρ values show the domain ordering survives the change of reference for every model on the slate.

differences of one to two F1 points, and without per-item confidence intervals we cannot separate a genuine reordering from noise.

Neither parameter count nor model recency guarantees superior recall. On the original ground truth, Llama-3.1 8B (36.32 F1) outperforms the larger Mistral-Nemo (32.56), a gap that persists on the augmented reference (27.84 against 24.18). Newer architectural generations do not reliably outperform their predecessors either: Gemma-2 9B (36.92) beats both Gemma-3 12B (35.35) and Gemma-4 12B (32.25) despite being an older and smaller model. This variability suggests that differences in pre-training data curation and post-training alignment protocols may exert a stronger influence on factual recall than parameter count or release date alone.

A second protocol effect. Reference completeness is not the only protocol choice that moves a factuality score without the model changing. Evaluating thinking configurations exposes another: when a model interleaves a reasoning trace with its answer, the evaluator must decide what counts as the answer, and that decision alone moves alias-

match F1 by between 3 and 22 points for the Qwen3.5 series—larger than the spread the same table attributes to model scale. We report those results in Appendix F rather than here, because they characterise decoding behaviour rather than factual recall; the point they contribute is that answer extraction belongs alongside reference completeness on the list of protocol choices a factuality benchmark must pin down before its numbers mean anything.

6 Discussion

MUSART shows that scalable, multi-domain ground truth can be built from categorical structure plus graph salience, and that source-grounded augmentation, read as a conservative and an optimistic reference, supports completeness-aware claims without requiring a manually audited ground truth. A more complete benchmark yields a more humbling assessment: measured precision rises while measured recall falls sharply, suggesting models know a smaller fraction of the available facts than standard evaluation implies.

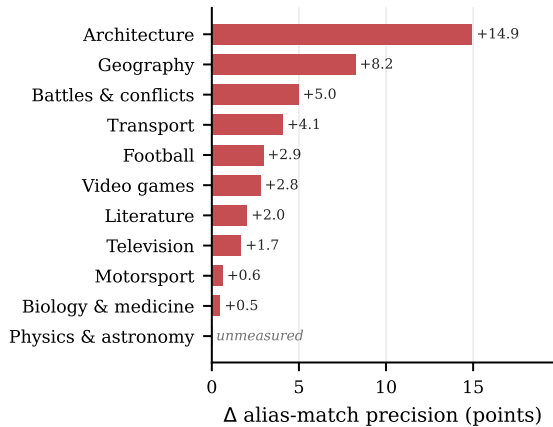


Figure 3: The correction gap by domain: augmented minus original alias-match precision, averaged over the twenty-one models. Wikidata’s incompleteness is not a constant that a single global adjustment could absorb—it is worth 14.9 points of measured precision in *Architecture* and 0.5 in *Biology and medicine*, a thirty-fold range. Crucially, this ranking measures knowledge-base deficiency, not domain difficulty: *Architecture* tops the chart because article text supplies objects models already produce but Wikidata lacked, whereas *Biology and medicine* sits at the bottom because its heavy augmentation adds objects the models do not know. *Physics and astronomy* is shown as unmeasured rather than as a zero: the extractor produced no output for that domain (Section 4).

In practice we recommend reporting against *both* ground truths, never the augmented one alone: a single augmented-GT figure launders extraction noise into apparent fact, whereas the pair exposes how far the score moves with the reference, and the size of that movement is itself informative—a wide gap flags domains or relations where Wikidata coverage is thin and completeness correction matters most. Cross-ground-truth comparisons should be made on precision; recall and F1 comparisons are meaningful only within one ground truth.

Furthermore, because all evaluated models share the same fundamental decoder-only Transformer architecture, their highly correlated domain profiles likely reflect shared structural properties and inductive biases rather than merely overlapping pre-training corpora. We suggest that future work could use mechanistic interpretability tools to investigate how the Transformer’s specific mechanisms for storing and retrieving relational facts give rise to this shared knowledge profile, and whether alternative architectures exhibit different factual recall patterns.

7 Conclusion

We presented a construction method for multi-domain factual benchmarks that needs no per-domain relation curation, MUSART as its instantiation on Wikipedia and Wikidata, a source-grounded augmentation pipeline whose paired references expose how far a factuality score moves with reference completeness, and a study of twenty-one open-weight LLMs that separates evaluation artefacts from genuine knowledge.

Three directions follow directly. A calibrated audit of the augmented ground truth would convert its reference gap from a sensitivity measurement into an estimate with guarantees. A second prompting condition would establish how much of what we report is specific to forced answering. And applying the construction method to a different corpus–knowledge-base pair would test the claim that its domain-discovery step is genuinely portable—the part of this work we expect to outlive the particular benchmark it produced.

Limitations

The augmented ground truth is un-audited under the pipeline configuration reported here (gemini-3.1-flash-lite extraction with hyper-link matching and string deduplication): it contains noise from extraction, and from unlinked single-surface-form objects that cannot be de-duplicated against Wikidata. Consequently, figures computed against the original and augmented ground truths should be read as two contrasting reference regimes rather than as point estimates, and a calibrated audit of the augmented ground truth is left to future work. Inferred (non-grounded) extractions are discarded, trading recall for precision. All results use a single prompting condition (few-shot), which narrows the generality of the findings; prompting sensitivity is left to future work. Domain-level estimates inherit the size of the domain: *Physics and astronomy* contributes only 113 questions and *Motorsport* 326, so their per-domain figures rest on thin samples and should not be over-read. MUSART is built from English Wikipedia and English question templates, so all findings concern English factual recall. Because Wikipedia is present in essentially every open-weight pre-training corpus, MUSART measures recall of likely-seen facts—the quantity of interest for the LM-as-KB setting—rather than generalisation to unseen knowledge. Finally, results derive from a fixed model set, and absolute num-

bers will change as newer open-weight LLMs are added.

Reproducibility

MUSART is regenerated end to end from public Wikipedia and Wikidata dumps by the pipeline of Section 3; the augmentation of Section 4 uses langextract with a pinned gemini-3.1-flash-lite version. All generation is greedy, so runs are deterministic given a fixed serving stack. All code, data, question templates, few-shot exemplars, scoring scripts, and the per-run summary statistics behind Table 3, Figure 2, and Figures 6 and 7 in Appendix C are publicly released at <https://anonymous.4open.science/r/MUSART/>.

Ethics Statement

MUSART is built from publicly available Wikipedia content (CC BY-SA 4.0) and Wikidata content (CC0), and will be released under a compatible license with attribution to the source projects. The benchmark concerns factual recall about encyclopaedic entities and poses no foreseeable harm beyond standard considerations for automatically constructed datasets.

References

- Thomas Bonald, Nathan de Lara, Quentin Lutz, and Bertrand Charpentier. 2020. [Scikit-network: Graph analysis in python](#). *Journal of Machine Learning Research*, 21(185):1–6.
- Xilun Chen, Iliia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025. [Learning to reason for factuality](#). *Preprint*, arXiv:2508.05618.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. 2026. [Noisy but valid: Robust statistical evaluation of llms with imperfect judges](#). *Preprint*, arXiv:2601.20913.
- Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2013. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 413–422.
- Yujia Hu, Shrestha Ghosh, Tuan-Phong Nguyen, and Simon Razniewski. 2024. [GPTKB: Building very large knowledge bases from language models](#). *Preprint*, arXiv:2411.04920.
- Shantanu Jain, Martha White, and Predrag Radivojac. 2016. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Jan-Christoph Kalo and Leandra Fichtel. 2022. KAMEL: Knowledge analysis with multitoken entities in language models. In *4th Conference on Automated Knowledge Base Construction (AKBC)*.
- 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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12076–12100.
- Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Rashika Ramola, Shantanu Jain, and Predrag Radivojac. 2019. Estimating classification accuracy in positive-unlabeled learning: Characterization and correction strategies. In *Pacific Symposium on Biocomputing (PSB)*, volume 24, pages 124–135.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling LLM test-time compute optimally can be more effective than scaling model parameters](#). *Preprint*, arXiv:2408.03314.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? a.k.a. will llms replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. *Measuring short-form factuality in large language models*. Preprint, arXiv:2411.04368.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24824–24837.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. *Are reasoning models more prone to hallucination?* Preprint, arXiv:2505.23646.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.

Appendix

A Construction Pipeline and Graph-Based Saliency Formulation

Figure 4 illustrates the automated MUSART benchmark construction pipeline, which seeds from Wikipedia Good Article categories and aligns to Wikidata without manual per-domain schema curation.

Granular Triple Acquisition and Filtering Steps.

To supplement the summary in Section 3.1, the initial triple collection proceeds as follows. First, we select articles designated as “Good Articles” (GAs) on the English Wikipedia, which constitute approximately 0.6% of total articles and satisfy explicit editorial criteria for factual accuracy, neutrality, verifiability, and stability. From the “Wikipedia:Good articles/all” page, we harvest 38,279 GA titles organ-

Domain	Salient Relations
Biology & med.	hymenium attachment, IUCN conservation status, mushroom ecological type, has basionym, has cause, health specialty, studied in, mushroom cap shape, stipe character, edibility, spore print color, subject has role, original combination, diel cycle, hymenium type
Geography	located in or next to body of water, mountain range, continent
Places	historic county, shares border with, located in time zone, located in or next to body of water, capital of, twinned administrative body
Transport	terminus location, state of use, connects with, country, connecting service, maintained by, transport network, connecting line, located in the administrative territorial entity, adjacent station, owned by, terminus

Table 4: Examples of manually selected salient relations for four Good Articles categories, illustrating the desired output of our automated methodology.

ised under 65 second-level domain categories. Second, using Wikimedia dump mapping tables, we link each GA title to its Wikidata QID; 11,353 articles lacking a QID are discarded to retain 26,926 linked entities, of which 26,210 yield statement triples in Wikidata. Third, from the Wikidata dump of truthy triples, we harvest 302,179 statement triples (s, r, o) across 1,001 relations where s is a linked GA QID. Finally, we apply two-stage filtering: (1) removing 24 generic, administrative, or Wikimedia-internal properties (e.g., instance of; see Table 6) and all ID-bearing properties, and (2) thresholding relation frequency at ≥ 50 occurrences across the corpus (removing 689 sparse relations whose median frequency is 13) so that subsequent saliency scoring operates on representative properties.

Table 4 presents examples of manually curated salient relations across four representative domains, illustrating the target domain-specificity that our automated graph-based saliency pipeline approximates.

Formal Mathematical Formulation of Graph-Based Saliency.

To formalise the Personalized PageRank (PPR) saliency selection introduced in Section 3, we construct a heterogeneous directed graph $G = (V, E)$ from the filtered Wikidata triples and their associated Wikipedia data. The node set V is the disjoint union of:

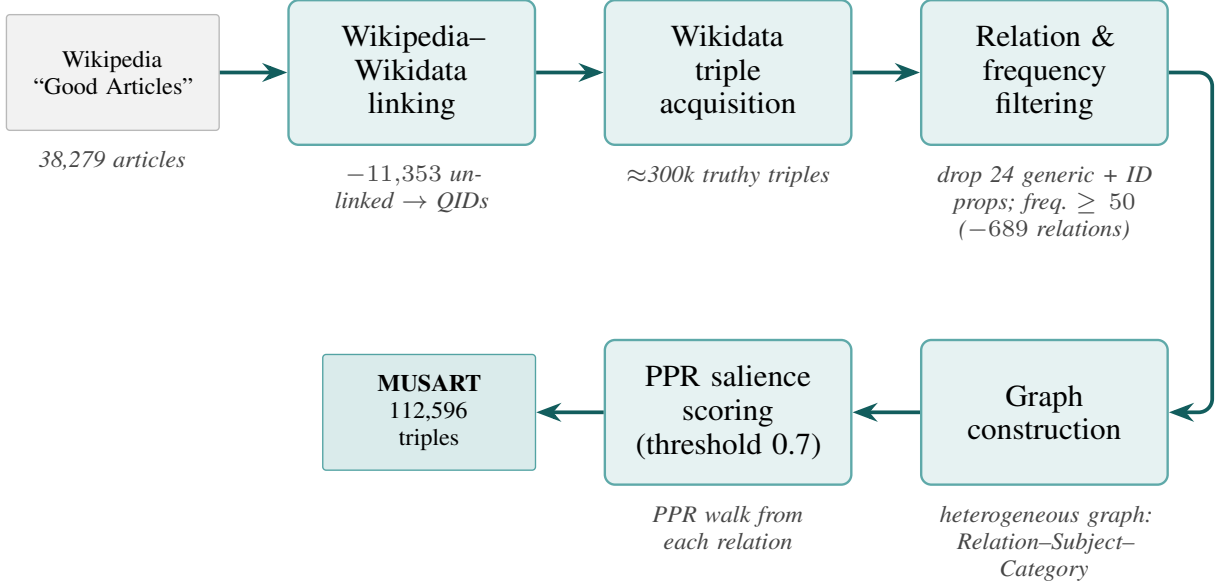


Figure 4: Construction of MUSART. An automated pipeline links high-quality Wikipedia “Good Articles” to Wikidata, acquires their triples, removes generic and rare relations, and scores the remainder with a Personalized PageRank salience model, distilling $\approx 300k$ raw triples and 1,001 relations down to 112,596 triples over 173 salient relations across 31 domains.

- Subject nodes V_S representing Wikidata subjects;
- Property nodes V_R representing Wikidata relations;
- Category nodes V_C corresponding to second-level Wikipedia categories extracted from the Good Article hierarchy.

Two typed edge sets are defined:

- $E_{SC} = \{(s, c) \mid s \in V_S, c \in V_C, s \text{ is labelled with } c\}$, linking each subject to its category;
- $E_{RS} = \{(r, s) \mid r \in V_R, s \in V_S, (s, r, \cdot) \text{ appears in the filtered triples}\}$, linking every relation to the subjects in which it occurs.

The complete edge set is $E = E_{SC} \cup E_{RS}$. We use the `scikit-network` library (Bonald et al., 2020) to represent the graph and compute the Personalized PageRank. We run PPR initiated from each unique relation node $r \in V_R$. The PageRank score obtained by a category node $c \in V_C$ during this random walk reflects the structural strength of the connection between relation r and category c . Sorting relations by this score yields a ranked list per category, where an empirical cutoff of 0.7 isolates salient domain relations.

B Ground-Truth Augmentation Pipeline Details and Domain Statistics

Figure 5 depicts the ground-truth augmentation pipeline introduced in Section 4, which enriches Wikidata triples with character-grounded facts extracted verbatim from Wikipedia Good Article texts.

Granular Extraction and Linking Methodology.

The augmentation pipeline operates via three transparent stages:

1. **Candidate Extraction:** For each subject-relation pair, we perform extractive question answering on the corresponding Good Article text using Gemini 3.1 Flash Lite via the `langextract` library.¹ A key feature of `langextract` is its **precise source grounding**, mapping every extraction to its exact character interval in the source document.
2. **Candidate Tiering:** Extractions are categorised into two tiers based on alignment metadata. **Tier 1 (Grounded)** candidates are those returning an `alignment_status` of `match_exact`, confirming that the extracted string is a direct, verbatim quote from the text. **Tier 2 (Inferred)** candidates lack exact character grounding and rely on model inference;

¹<https://github.com/google/langextract>

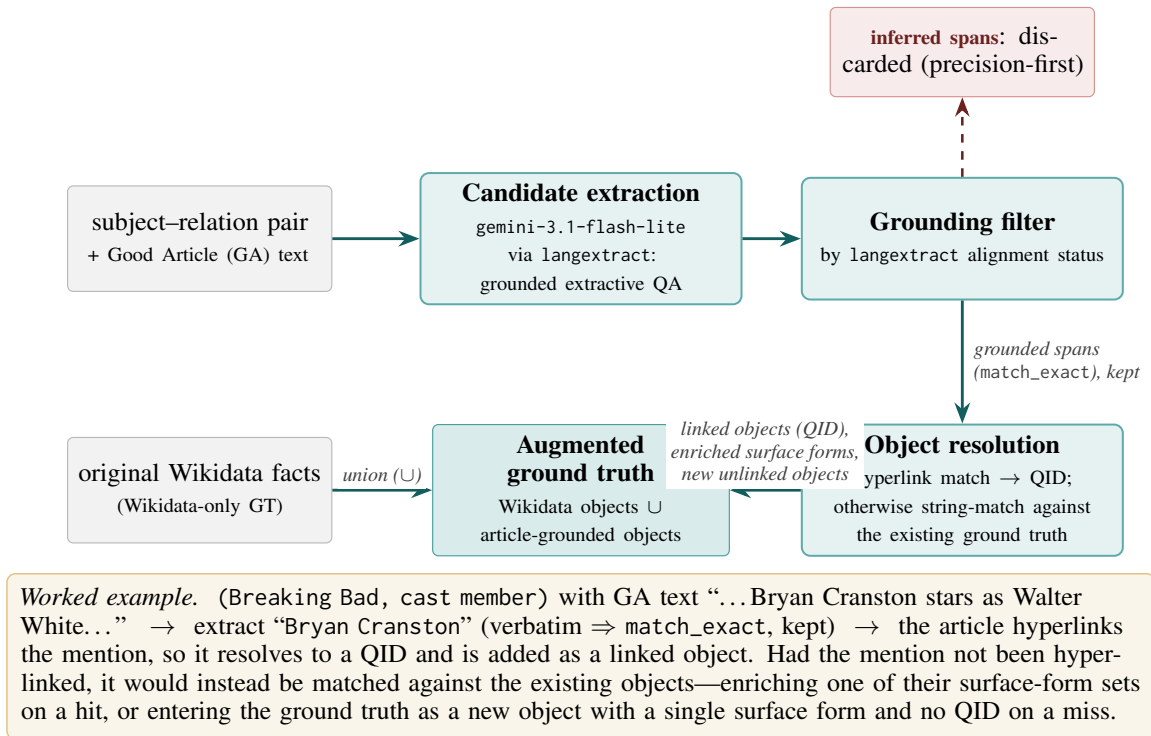


Figure 5: Augmenting the MUSART ground truth. An automated, precision-first pipeline extracts candidate objects from “Good Article” text, keeps only verbatim-grounded candidates, and resolves them against Wikidata. Hyperlink matching is the only step that assigns QIDs; remaining grounded spans are string-matched against the existing ground truth, either enriching the surface forms of an object already present or entering as a new object carrying one surface form and no QID. The result is the augmented ground truth used throughout Section 5; it is not audited, and the consequences of that are set out in Section 4.

to prioritise precision, Tier 2 extractions are discarded.

3. **Entity Linking and Deduplication:** To resolve extracted spans to canonical Wikidata entity IDs (QIDs), we rely exclusively on hyperlink matching. We parse the source article HTML to determine if an extracted span coincides with an anchor hyperlink; if hyperlinked, we map the target page title to its QID via Wikimedia dumps. Unlinked extractions are retained as surface forms without external parametric linking models. To avoid duplicates, unlinked strings are matched against existing Wikidata labels and aliases for that subject–relation pair: matching strings are merged into the existing Wikidata entity, while non-matching strings are retained as newly discovered, unlinked object entities.

Domain Distributions and Evaluation Properties of Unlinked Objects. As reported in Table 5, augmentation mass varies widely across domains: it multiplies the gold-object count by $4.8\times$ in *Battles, exercises, and conflicts* but by $1.14\times$ in *Motorsport*, with linking rates ranging from 70% (*Geography*) to 16% (*Biology and medicine*). Ratio and absolute mass diverge: *Video game titles* has a modest $1.8\times$ multiplier but contributes 14,157 of the 45,205 objects added by augmentation. Extraction ran at full coverage on ten of the eleven evaluated domains; *Physics and astronomy* (113 questions, drawn entirely from epoch) produced no extraction output and is treated as unmeasured rather than measured-as-complete in Section 5.3. Furthermore, two properties of unlinked objects affect evaluation: (1) an unlinked object carrying a single surface form and no alias set can only be matched by that exact string, depressing mean surface forms per object (Table 5); and (2) lacking a QID, an unlinked object cannot be de-duplicated against existing Wikidata entities, which can slightly inflate object counts and depress measured recall while introducing a minor false-positive risk for noisy spans.

Table 5 reports the per-domain impact of this augmentation pipeline across all eleven evaluated MUSART domains.

C Additional Evaluation Figures

This appendix provides supplementary visual analyses of model evaluation across reference ground truths and subject domains.

Domain	Objects/question			
	orig.	aug.	$\times$	% link
Battles & conflicts	2.52	12.01	4.8	37.9
Transport	1.42	3.88	2.7	68.1
Geography	1.18	3.17	2.7	70.0
Architecture	1.22	3.08	2.5	66.3
Biology & medicine	1.54	3.51	2.3	16.1
Literature	1.14	2.23	1.9	43.7
Video game titles	2.18	3.91	1.8	41.5
Television	3.36	5.77	1.7	61.3
Football	1.55	2.24	1.4	53.9
Motorsport	1.00	1.14	1.1	61.7
Physics & astronomy	1.00	1.00	1.0	—

Table 5: Augmentation mass by domain: gold objects per question before and after augmentation, the resulting multiplier, and the share of extracted mentions resolved to QIDs by hyperlink matching. *Physics and astronomy* yielded no retained extractions.

Figure 6 illustrates the reference sensitivity of measured alias-match precision for each evaluated open-weight LLM, comparing scores against the conservative Wikidata-only reference and the optimistic augmented reference.

Figure 7 displays per-domain precision and recall under both ground truths, illustrating where Wikidata is incomplete relative to model outputs (the precision gap) versus how much article text adds that models do not recall (the recall drop).

D Filtered Generic Relations

Table 6 lists the 24 relations excluded from our analysis. These encode Wikidata/Wikimedia bookkeeping or organisational structure (e.g. instance of, subclass of, topic’s main category) rather than factual statements about the entity itself, and therefore do not constitute meaningful targets for factual-knowledge evaluation.

E Salient Relations by Domain

Below, we list the relations identified as salient by the PPR selection procedure (Section 3.2) for a broader intermediate set of 50 categories processed during benchmark construction. (As noted in Section 3.1, manual selection and cardinality filtering yielded the final 11 domains included in the released dataset). Categories for which no relation passed the salience threshold are listed at the end.

Agriculture, food, and drink. this taxon is source of, product, material, or service produced or provided, taxon range

Albums. tracklist, place of publication, charted in, recorded at studio or venue, record label, form of creative work, performer, producer, language of

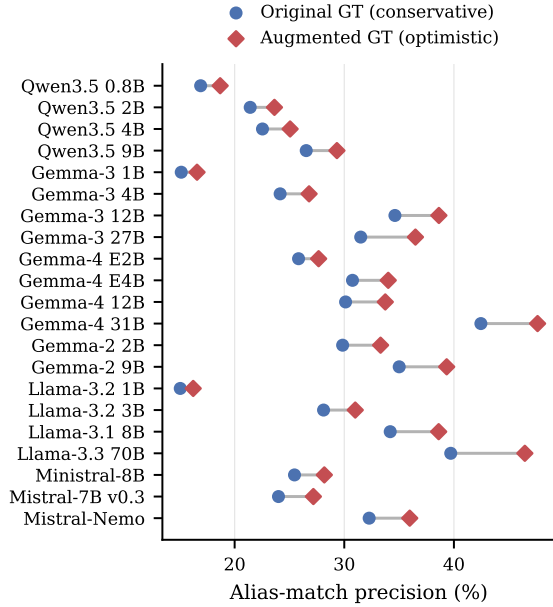


Figure 6: Reference sensitivity of measured precision. For each model, alias-match precision against the original (Wikidata-only) ground truth—a conservative reference—and against the unfiltered augmented ground truth—an optimistic one. The connecting segment is how far the score moves when only the reference changes; it is neither an improvement nor an interval containing the true value.

work or name, distribution format, cover art by, has characteristic, genre, followed by, follows

Architecture. diocese, located on street, architectural style, architect, dedicated to, occupant, heritage designation, has facility, commissioned by, made from material, crosses, historic county, located in the administrative territorial entity, has use

Armies and military units. military branch, parent organization, has subsidiary, allegiance, item operated

Art. collection, depicts, movement, made from material, creator, copyright status as a creator

Battles, exercises, and conflicts. participant, location, part of

Biology and medicine. health specialty, diel cycle, original combination, CITES Appendix, hymenium type, mushroom cap shape, hymenium attachment, stipe character, spore print color, mushroom ecological type, edibility, NZTCS conservation status, IUCN conservation status, endemic to, taxonomic type, has basionym, parent taxon, taxon rank, taxon synonym, taxon range, this taxon is source of, has cause, location of discovery, studied in, subject has role, has effect, opposite of, practiced by, has use, described by source, said to be the same as, uses

Chemistry and materials science. location of discovery, discoverer or inventor, has effect, subject has role

Compositions. movement, has characteristic, tonality

Computing and engineering. copyright license,

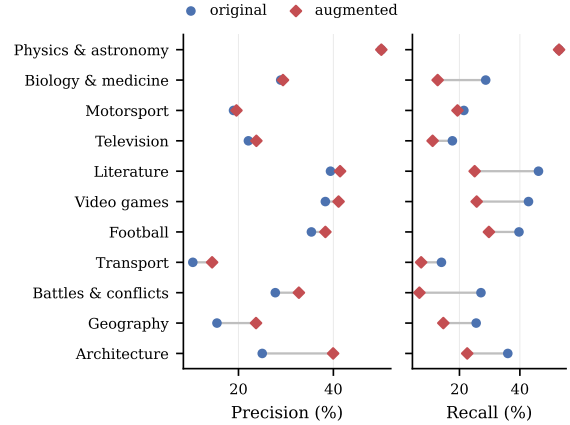


Figure 7: Per-domain alias-match precision and recall under both ground truths, averaged over the twenty-one models and ordered by the size of the precision gap. That gap measures where Wikidata is incomplete relative to what models produce; the recall drop measures how much the article text adds that models do not produce. The two orderings differ.

operating system, product, material, or service produced or provided

Cue sports. sponsor, organizer, winner, sports season of league or competition

Earth science. located in/on physical feature, time period, part of the series

Economics and business. industry, product, material, or service produced or provided, legal form, headquarters location, founded by, has subsidiary

Education. category for alumni of educational institution, has subsidiary, language used, member of, parent organization

Film. RTC film rating, ICAA rating, CNC film rating (Romania), CNC film rating (France), EIRIN film rating, MPA film rating, Medierådet rating, JMK film rating, production designer, costume designer, RCQ classification, assessment, Filmiroda rating, FSK film rating, film editor, after a work by, director of photography, IFCO rating, ClassInd rating, NMHH film rating, Kijkwijzer rating, filming location, BBFC rating, IMDA rating, production company, fabrication method, executive producer, soundtrack release, cast member, color, nominated for, screenwriter, set in period, director, original language of film or TV show, composer, based on, narrative location, derivative work, main subject, distributed by, Australian Classification, producer, award received, inspired by, set during recurring event, set in environment, collection, said to be the same as, copyright status, has edition or translation, country of origin, distribution format, characters, genre, voice actor, depicts

Football. referee, head coach, season of club or team, home venue, league or competition, category for members of a team, participating team, competition class, sports season of league or competition, sport, organizer, time period, chairperson, occupant, winner, is a list of, owner of, history of topic, headquarters location

Relation
instance of
topic's main category
subclass of
KML file
sex or gender
on focus list of Wikimedia project
topic's main template
aspect ratio (W:H)
religion or worldview
facet of
maintained by WikiProject
category of associated people
category for people born here
category for people who died here
category for maps or plans
permanent duplicated item
topic's main Wikimedia portal
has part(s) of the class
category for the interior of the item
ethnic group
member category
category for the view from the item
FactGrid item ID
category for the view of the item

Table 6: The 24 excluded generic relations.

Geography. tributary, volcano observatory, mouth of the watercourse, mountain range, drainage basin, highest point, located in protected area, continent, located in or next to body of water, made from material, located in the administrative territorial entity, historic county

Law. court, applies to jurisdiction, published in

Literature. intended public, illustrator, author, has edition or translation, published in, cover art by, enemy, from narrative universe, derivative work, student of, main subject, present in work, hair color, writing language, copyright status as a creator, publisher, list of characters, copyright status, language of work or name, eye color, inspired by, creator, narrative role, notable work, owner of, form of creative work, relative, spouse, residence, native language, child, sibling, occupation, has list, family name, languages spoken, written or signed, country of origin, military or police rank, field of work

Mathematics and mathematicians. studied in

Media and drama. eye color, narrative role, hair color, unmarried partner, inspired by, relative, residence

Military aircraft. powered by, armament

Military people. military or police rank, allegiance

Motorsport. pole position, fastest lap, winner, uses, sport, organizer

Multi-sport event. flag bearer, participant in, sport, part of the series

Other music articles. discography, instrument, location of formation, presenter, influenced by, destination point, notable work, educated at, start point, native language, writing language, has list, languages spoken, written or signed, family name, based on, participant in, field of work, has part(s), award received

Other sports. destination point, winner, start point, competition class, participating team, has list, practiced by, sport

Philosophy. student of

Physics and astronomy. parent astronomical body, discovery method, constellation, epoch, discoverer or inventor, start point, crew member

Places. twinned administrative body, head of government, office held by head of government, capital of, shares border with, authority, located in time zone, official language, capital, language used, located in or next to body of water, highest point, history of topic, owner of, historic county, continent, replaces

Politics and government. office contested, chairperson, applies to jurisdiction, member of political party

Religion. canonization status, office contested, student of, writing language, position held, place of death, copyright status as a creator, spouse, mother, place of burial, father, occupation, unmarried partner, place of birth

Royalty, nobility, and heraldry. family, noble title, place of burial, position held, mother, sibling, spouse, father, child, member of political party, place of death, cause of death, manner of death, canonization status, relative, given name, country of citizenship, unmarried partner, place of birth, occupation, depicts, applies to jurisdiction, military or police rank, depicted by, languages spoken, written or signed, copyright status as a creator, time period, native language

Songs. lyricist, tonality, performer, record label, form of creative work, producer, charted in, recorded at studio or venue, composer, part of, genre, follows, followed by, language of work or name, published in, has characteristic

Television. season, original broadcaster, list of episodes, theme music, original language of film or TV show, screenwriter, voice actor, set during recurring event, media franchise, director, takes place in fictional universe, characters, part of the series, color, narrative location, presenter, set in period, cast member, executive producer, set in environment, distributed by, follows, followed by, list of characters, production company, Kijkwijzer rating, filming location, creator, first appearance, country of origin, has part(s), distribution format, NMHH film rating, inspired by, fabrication method

Transport. terminus location, fare zone, transport network, adjacent station, connecting line, connecting service, terminus, track gauge, located on linear feature, state of use, maintained by, connects with, crosses, owned by, item operated, partially coincident with, located in the administrative territorial entity, destination point, replaces, start point, depicted by, country, powered by, operator, replaced by, is a list of, historic county, has cause, located in protected area, has facility

Video game characters, concepts, and miscellanea. first appearance, narrative role, from narrative universe, present in work, hair color, enemy, subject has role, eye color, product, material, or service produced or provided, legal form, industry, parent organization, residence, creator, family name, given name, armament, is a list of

Video game titles. game mode, CERO rating, PEGI rating, USK rating, compatible with, ESRB rating, software engine, input device, platform, developer, operating system, copyright license, Australian Classification, publisher, designed by, list of characters, distributed by, set in environment, composer, distribution format, soundtrack release, takes place in fictional universe, nominated for, IMDA rating, characters, narrative location, set in period, country of origin, part of the series, genre, BBFC rating, derivative work, copyright status, media franchise, language of work or name, based on, award received, set during recurring event, has characteristic, voice actor, uses

Warships and naval units. vessel class, country of registry, shipping port, location of creation, crew member, conflict, manufacturer, operator, armament, significant event, named after, powered by, different from, allegiance, described by source

Weapons, equipment, and buildings. commissioned by, has facility, historic county, heritage designation, depicted by

World history. capital, official language, replaced by, language used, replaces, allegiance, continent, has cause

Domains without salient relations identified. Baseball; Basketball; Cricket; Hockey; Language; Magazines and print journalism; Military decorations and memorials; Pro wrestling; Recreation.

F Thinking Configurations

Eight of the twenty-one models expose a thinking mode: all four Qwen3.5 sizes and all four Gemma-4 variants. We evaluated all of them, and report the results here rather than in Section 5 because they turn out to characterise decoding behaviour rather than factual recall, and because the two families require different generation protocols, which makes a single side-by-side column misleading.

Protocol. Gemma-4 thinking runs use `max_new_tokens = 1,024` with no separate reasoning budget. Qwen3.5 thinking runs cap the reasoning trace at 1,024 tokens—vLLM forces the closing tag at that point—within a total budget of 4,096 tokens. The Qwen configuration therefore changes two things at once relative to the earlier uncapped run: it introduces the reasoning cap *and* quadruples the total budget. The observed change in the closing-tag rate, from 0% to 100%, is almost certainly the cap rather than the budget, but this run does not formally isolate the two. Thinking traces are stripped by the scorer before matching; all figures below are computed over the full 21,775-item set unless stated otherwise.

Qwen3.5: a decoding failure, not a reasoning deficit. Table 7 gives the picture. Over the full item set, thinking is a net loss for every Qwen3.5

		AM F1, all items		clean
Model	% clean	direct	thinking	Δ
<i>Original ground truth</i>				
Qwen3.5 0.8B	42.3	11.65	3.99	+2.56
Qwen3.5 2B	81.8	17.89	11.38	+4.92
Qwen3.5 4B	56.1	26.18	4.12	+4.85
Qwen3.5 9B	85.3	30.53	13.76	+7.07
<i>Augmented ground truth</i>				
Qwen3.5 0.8B	42.3	7.74	3.33	+1.77
Qwen3.5 2B	81.8	12.40	9.53	+3.21
Qwen3.5 4B	56.1	20.42	4.28	+2.18
Qwen3.5 9B	85.3	23.57	13.04	+4.23

Table 7: Qwen3.5 thinking configurations. “% clean” is the share of items on which the model emits a usable answer after the trace is stripped; “direct” and “thinking” are alias-match F1 over all 21,775 items; Δ is the thinking-minus-direct difference *restricted to the clean subset*. Thinking loses over the full set and gains on the subset where it answers.

size—between 3 and 22 alias-match F1 points below the same model answering directly, and above 6 points in five of the six model–reference combinations. But on the subset of items where the model emits a usable answer, thinking is a consistent *gain* of 1.8 to 7.1 points, in every size and against both ground truths. The two facts are reconciled by the third column: the model fails to produce any usable answer on 15–58% of items, entering a repetition loop that consumes the budget. Median answer length after stripping is 22–91 characters for the 2B, 4B and 9B models, but 9,585 characters for the 0.8B, which essentially never terminates. In the earlier uncapped runs the trace was never closed and 75–90% of items were empty after stripping, driving alias-match F1 down to 0.4–6.9; those numbers measure a chat-template interaction rather than model capability, and we do not report them as such.

The clean-subset figures are not a fair headline—they are computed on a self-selected population and are not comparable to any other number in this paper—but they do establish that the deficit is in terminating the answer, not in recalling the fact.

Gemma-4 behaves differently. The failure is Qwen-specific, not a property of thinking runs in general. Under the scorer’s own stripping path, Gemma-4 thinking runs produce short, well-formed answers—median 18–28 characters—and essentially no degeneration: at most 0.7% of answers exceed 200 characters, against 9–23% for the uncapped Qwen runs. Their empty-after-stripping

rates are 3.8% (E4B), 6.6% (E2B), 12.5% (31B) and 33.8% (12B). For gemma-4-12B-it, the model common to both analyses, thinking raises alias-match F1 from 32.25 to 33.45 against the original ground truth and leaves it essentially unchanged against the augmented one (24.58 to 23.10)—a far smaller effect than the Qwen degeneration, and one that does not survive the change of reference. Its 33.8% empty rate means even this figure is depressed, for a different reason than Qwen’s: truncation rather than looping.

What this contributes. The relevance to MUSART is methodological. For a single unchanged model, Qwen3.5-4B, alias-match F1 against the original ground truth is 4.0 under the uncapped protocol, 4.1 under the capped protocol scored over all items, and 35.6 under the capped protocol scored over answered items. Three defensible-sounding choices about what counts as the model’s answer produce a ninefold spread, without the model or the benchmark changing. The answer-extraction rule is therefore a protocol variable of the same order as the completeness of the reference—the effect quantified in Section 5.2—and a factuality benchmark that leaves it implicit is not reporting a reproducible number. This is the sense in which the thinking axis supports the paper’s argument: not as a claim about reasoning, but as a third demonstration that protocol choices move factuality scores as much as models do.

G Evaluation Setup Details and Prompt Format

From triples to questions. Following the LM-as-KB paradigm (Petroni et al., 2019), we convert each (subject, relation) pair into a natural-language question using relation-specific templates: for example, for heritage designation the template is “What is [S]’s heritage designation?”, where [S] is the subject label. Templates were drafted with an LLM and then manually reviewed and corrected for accuracy and naturalness. Prompts are generated from the structured triples alone—the article text is never shown to the evaluated model—so answers reflect parametric knowledge only. Because many questions admit several valid objects, models are instructed to list each answer on a separate line, and evaluation is set-based.

Prompting and scoring mechanics. Each model answers under five-shot prompting: every test ques-

tion is preceded by five in-context question–answer exemplars drawn from the same relation and subject category, teaching the expected answer type, surface form, and one-answer-per-line format; exemplar questions are excluded from the evaluation set to prevent leakage. All models are served with vLLM under greedy decoding (temperature 0), with a generation budget of 50 new tokens.

We score model responses with set-based *alias matching* (AM). Every metric (precision, recall, and F1) is reported as a percentage ($\times 100\%$). Each gold object o in the reference set O is defined by a canonical Wikidata label together with an alias set $\mathcal{A}(o) = \{\ell_1, \dots, \ell_k\}$ of known surface forms, including Wikidata aliases and Wikipedia hyperlinked anchor texts (mean 4.29 surface forms per object, up to 7.82 in *Battles, exercises, and conflicts*). Given a model’s line-separated predictions $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_m\}$, each string $\hat{y}$ is normalised by lowercasing and trimming punctuation. A prediction $\hat{y}$ matches gold object o ($\text{Match}(\hat{y}, o) = 1$) if and only if its normalised form is an *exact match* to at least one normalised alias in $\mathcal{A}(o)$:

$$\text{Match}(\hat{y}, o) = \mathbb{I}[\exists \ell \in \mathcal{A}(o) : \text{norm}(\hat{y}) = \text{norm}(\ell)]. \quad (1)$$

Rejection of substring false positives. Crucially, alias matching requires full string identity against an alias, *not* substring containment. For example, if the gold entity is *New York City Subway* (whose alias set includes *NYC Subway* and *New York Subway*), returned answers such as “City” or “New” are strictly false positives (0 credit), preventing generic fragments from inflating scores. A generic term such as “Metro” is credited if and only if it is explicitly registered in that specific entity’s alias set; otherwise, it receives zero credit.

Each prediction $\hat{y} \in \hat{Y}$ can credit at most one gold object, and each gold object $o \in O$ is credited at most once. Precision is the fraction of predicted answers that match a gold object ($|Y_{\text{correct}}|/|\hat{Y}|$), recall is the fraction of gold objects matched ($|O_{\text{recalled}}|/|O|$), and F1 is their harmonic mean, micro-averaged over all evaluation questions.

Model slate design. Across our twenty-one evaluated open-weight models (0.8B to 70B parameters), *scale* varies within a family and generation: Qwen3.5 at 0.8B/2B/4B/9B, Gemma-3 at 1B/4B/12B/27B, Gemma-4 at E2B/E4B/12B/31B, and the Llama-3 series at 1B, 3B, 8B and 70B. *Generation* varies at roughly matched size: Gemma-2 at 2B and 9B against Gemma-3 and Gemma-4 at comparable scales. *Family* varies at matched scale in the 7–12B band, where Mistral-7B-

You are an AI model designed to answer questions in the following format: “What transport network is the infrastructure [S] part of?”. Guidelines for Your Response: - [S] is the subject provided by the user and belongs to the category “Transport”. - Only include the relevant answer object(s)—without any extra text, explanations, or formatting. - If multiple objects exist, list each one on a separate line. Examples of questions and answers: What transport network is the infrastructure Westminster tube station part of? London Underground What transport network is the infrastructure Gare de Lyon part of? Paris Métro Réseau Express Régional What transport network is the infrastructure Alexanderplatz station part of? Berlin U-Bahn Berlin S-Bahn What transport network is the infrastructure Shibuya station part of? Tokyo Metro Yamanote Line What transport network is the infrastructure Flinders Street railway station part of? Melbourne rail network What transport network is the infrastructure King’s Cross St. Pancras tube station part of?

Figure 8: Few-shot prompt (no abstain instruction) for the relation `part of`, Transport domain. In-context exemplars are drawn from the same relation and subject category as the test question.

v0.3, Ministral-8B and Mistral-Nemo sit alongside Llama-3.1-8B, gemma-2-9b-it and Qwen3.5-9B. Every score in this paper is computed by a single scoring pipeline over the stored generations of all twenty-one models against both ground truths.

Prompt format. Figure 8 shows the few-shot prompt format used in all evaluations, instantiated for the relation `part of` in the Transport domain. The five in-context exemplars share the test question’s relation and subject category and are excluded from the evaluation set. A variant of this prompt adds an “If you don’t know the answer, respond with ‘I don’t know’” guideline; the single condition reported here omits it, so every question is answered.