A Wikidata-Based Framework to Measure Cross-Lingual Bias in Multilingual Large Language Models

Mouloud Iferroudjene[1], Lisa Poggel[2], Andrea Schimmenti[3], Kanchan Shivashankar[4], Duo Yang[5], Jan-Christoph Kalo[6], Marta Boscariol[7]
[1]Mines Saint-Étienne, France, [2]Freie Universität Berlin, Germany, [3]Università di Bologna, Italy
[4]Bergische Universität Wuppertal, Germany, [5]KU Leuven, Belgium
[6]University of Amsterdam, The Netherlands, [7]Università degli Studi di Torino, Italy
mouloud.iferroudjene@emse.fr, l.poggel@fu-berlin.de, andrea.schimmenti2@unibo.it,
shivashankar@uni-wuppertal.de, duo.yang@kuleuven.be, j.c.kalo@uva.nl, marta.boscariol@unito.it

Abstract

Multilingual large language models (LLMs) are increasingly used for factual question answering, yet their accuracy varies across languages in ways that are difficult to interpret. A central challenge is that many multilingual probing benchmarks conflate multiple factors: the language used to ask the question, the cultural-linguistic context of the entities being queried, and the popularity skew of entities. In our paper, we disentangle these factors by asking: (i) how strongly does the Language of the Question (LoQ) affect factual recall, (ii) does matching LoQ to an entity-associated Language of the Entity (LoE) improve performance, and (iii) do these effects persist when entity popularity is controlled. To this end, we introduce WILA-PopQA, a new Wikidata-grounded benchmark spanning 9 languages with matched popularity profiles, and probe 12 open-weight models of varying sizes and architectures under aligned and misaligned LoQ–LoE conditions. We evaluate models' answers to 4 types of questions about entity biographical properties in all selected languages. Results show that LoQ is the dominant source of variation. LoQ–LoE alignment does not consistently yield the highest accuracy, and performance depends on the property being asked. These results suggest that prompt language is an actionable experimental factor for multilingual factual evaluation.

Keywords: Multilingual LLMs, Wikidata, Factual Probing, Popularity Bias, Cross-Lingual Evaluation

1. Introduction

Large Language Models (LLMs) serve as interfaces to factual knowledge for open-domain question answering, knowledge retrieval, and Knowledge Graph (KG) construction. Their reliability, however, varies by language: a model may answer accurately in one language but fail or alter details in another. Prior multilingual factual probing datasets (such as X-FACTR (Jiang et al., 2020) and BMLAMA (Qi et al., 2023a)) document significant cross-language variation in factual recall, yet they tend to focus on a fixed set of "universal" facts, favor globally prominent entities, and conflate prompt-language effects with entity selection bias. This raises a question relevant to KG+LLM workflows: when users ask about entities tied to specific linguistic or cultural contexts, does the query language affect answer accuracy, and can such effects be measured in a reproducible, controlled way?

Addressing this question requires disentangling three factors that multilingual evaluations often confound. The Language of the Question (LoQ) changes the prompt's surface form and can shift model behavior. The Language of the Entity (LoE) captures the language in which an entity is primarily described or culturally situated, thereby affecting the availability of evidence in pre-training data. Entity popularity is an additional confound: Wikidata/Wikipedia-derived probes often over-sample globally visible entities, which tend to have denser metadata and broader label coverage. Popularity has been used as a quantitative signal in KG+LM probing (Arnaout et al., 2022), and sampling artifacts in Wikidata-based datasets have also been reported (Wiland et al., 2024).

We propose a Wikidata-based framework to probe multilingual LLMs on culturally grounded factual questions, making explicit the relationships among LoQ, LoE, and entity popularity within a single evaluation pipeline. Wikidata is well-suited to this purpose, as it provides a cross-lingual KG with multilingual labels, structured ground truth, and Wikipedia-linked sitelinks to operationalize popularity. We select 9 languages (**Arabic, Chinese, English, French, German, Hindi, Italian, Polish, and Russian**) spanning different language families, scripts, and levels of coverage in Wikidata. Entity sampling applies sitelink-equidistribution constraints to ensure that observed language effects are not reducible to differential entity prominence across language groups. Our study aims to answer the following research questions:

- **RQ1.** How does LoQ affect factual accuracy in multilingual LLMs for entities tied to specific linguistic and cultural contexts?
- **RQ2.** Can our Wikidata-based methodology

Proceedings of the Knowledge Graphs and Large Language Models Workshop (KG-LLM) @ LREC26, pages 190–209
May 16, 2026.

isolate language effects while accounting for LoE and differences in data availability?

- **RQ3.** Which quantifiable factors help explain performance differences after controlling for LoQ and LoE?

Contributions. This paper makes three contributions. First, we introduce a multilingual probing framework built from Wikidata that captures LoQ, LoE, and popularity in a single workflow. The framework generates prompts and multilingual ground-truth labels when Wikidata provides sufficient coverage. Second, we release WILA-PopQA, a new, curated Wikidata-derived benchmark covering 4 properties across 9 languages, with 2079 entities distributed equally in popularity. Third, we define an evaluation protocol for generative LLMs that reports standard task metrics and stratifies results by popularity to offer fair comparisons.

The rest of the paper is organized as follows. Section 2 reviews prior work on multilingual knowledge probing and cultural bias evaluation. Section 3 describes the dataset construction process and the collection of ground truth for multilingual labels. Section 4 presents the prompting setup and evaluation protocol. Section 5 reports the experimental results. Finally, Section 6 discusses the findings with respect to the research questions, highlights limitations, and outlines future work.

2. Related Work

Multilingual Knowledge Probing. The probing paradigm originates with Petroni et al. (2019), who transformed KG triples into cloze prompts to test whether masked language models could recover missing entities. Kassner et al. (2021) extended this to 53 languages via mBERT, reporting language-dependent variation in factual recall (FC). BMLAMA (Qi et al., 2023a) and DLAMA (Keleg and Magdy, 2023) introduced regionally stratified benchmarks across Western, Asian, and South American cultural contexts. Vu et al. (2024) reports performance degradation in medium- and low-resource languages using FActScore, attributing it in part to differential Wikipedia coverage. A recurring limitation across these datasets is the focus on universally prominent entities and the use of template or machine-translated prompts, which may conflate prompt-quality effects with language-level differences in FC metric (Youssef et al., 2023).

Cultural Knowledge and Bias in LLMs. Prior work has evaluated whether LLMs encode cultural knowledge uniformly across languages and regions. FORK (Palta and Rudinger, 2023) and CULTURE-GEN (Li et al., 2024) both find systematic divergence between Western and non-Western cultural contexts, with models performing more accurately on US-anchored questions. Naous et al. (2024) documents analogous asymmetries in Arabic versus Western knowledge representation, while Buyl et al. (2024) shows that models reflect the value orientations of their developers. Liu et al. (2024) proposes a taxonomy distinguishing cultural from sociocultural elements as a framework for structured evaluation. AlKhamissi et al. (2024) operationalize cultural alignment against demographically controlled survey data, finding that both pretraining language composition and prompt language influence alignment quality. MBBQ (Neplenbroek et al., 2024) further shows that bias levels vary significantly across languages within the same model.

Popularity. (Sun et al., 2024) measure popularity in DBpedia knowledge graph using traffic and density of entities and observes that factual retrieval correlates to the entity's popularity. Stronger models, such as GPT-4, struggle with long-tail knowledge. In another ranking task for long-tail cultural concepts, (Jiang and Joshi, 2024) found that close models such as GPT-3.5 perform better. In downstream behavior, such as recommendation generation, LLMs are a better substitute to traditional filtering systems, exhibiting fairness and mitigating popularity bias (Lichtenberg et al., 2024). Lastly, (Ni et al., 2025) examines LLMs' ability to assess their own knowledge boundary and finds that popular knowledge improves confidence and boosts correctness by more than 5%. Thus, establishing consistent evidence of close ties between LLM's ability to recall and reason over factual knowledge and its popularity.

Prompting Strategy and Language Effects. The choice of prompt language constitutes a distinct variable in multilingual factual retrieval. Wang et al. (2025) report a 14% reduction in hallucination rate when cultural cues and prompt language are aligned, while Rystrøm et al. (2025) find that the gap between language fluency and cultural alignment is not monotonically related to multilingual capability. Xu et al. (2023) proposes Language Representation Projection modules to improve cross-lingual factual transfer, and Qi et al. (2023b) shows that factual knowledge propagation across languages operates primarily through shared embeddings. The survey by Sahoo et al. (2024) documents the range of prompt engineering techniques that affect the quality of factual output.

Multilingual Resources and Coverage Limitations. Several resources address multilingual evaluation at the data level but target adjacent tasks. DaMuEL (Kubeša and Straka, 2023) provides a large-scale entity-linking dataset across 53

languages but does not support factual completion or KG probing. WDProp (Samuel, 2021) reports translation statistics for Wikidata property labels at the ontology level without enabling downstream factual evaluation. A persistent limitation across these resources is the uneven availability of multilingual entity labels in Wikidata: evaluation instances whose gold answers lack labels in a given language are typically either dropped—inflating the representation of globally prominent entities—or evaluated against English labels under non-English prompts, conflating LoQ effects with coverage artifacts (Arnaout et al., 2022).

3. Resources

3.1. WILA-PopQA: Popularity-matched multilingual Wikidata QA resource

Our dataset, WILA-PopQA [1], is based on Wikidata. The data comes from a local snapshot using the truthy dump (28.10.2023) [2], which retains only the highest-ranked statements for each subject–predicate pair.

Dataset Collection We select human entities (`wd:Q5`) using two criteria: language and occupation. To operationalize LoE, an entity must be linked to at least one of nine target languages through Wikidata `wdt:P1412` (*languages spoken, written, or signed*): English, French, German, Russian, Italian, Arabic, Polish, Chinese, and Hindi. This language set reflects the linguistic coverage of the author team and supports direct, qualitative validation of datasets. We select `wdt:P1412` to represent LoE rather than `wdt:P103` (*native language*) to balance semantic specificity and property coverage. Although `wdt:P103` more narrowly captures native language, it does not necessarily reflect the primary working language of an entity (e.g., writers whose publication language differs from their spoken native language). In addition, `wdt:P1412` is linked to considerably more item pages (3,447,069) than `wdt:P103` (372,597), resulting in a larger candidate set (Wikidata contributors, 2026). Each language is mapped to a defined set of Wikidata language items (QIDs) covering major variants and dialects (see Table 1).

For the occupation criterion, entities must have at least one occupation (`wdt:P106`) belonging to the following set: creator (`wd:Q2500638`), politician (`wd:Q82955`), actor (`wd:Q33999`), or writer (`wd:Q36180`). These occupations correspond to public and cultural figures with substantial web representation, making them suitable for cross-lingual factual probing.

After defining selection criteria, we construct the dataset following a three-step pipeline:

1. **Entity retrieval and property-presence profiling.** We retrieve entities matching our two criteria and compute a binary property-presence profile for each entity.

2. **Property selection and coverage scoring** We analyze cross-lingual property coverage to identify an ideal property subset that balances (i) the number of retained properties and (ii) the number of entities for which these properties are available across languages.

3. **Property value retrieval.** For the selected property set and the entities retained across languages, we retrieve the corresponding property values in all target languages (when available) to form the dataset.

Entity retrieval and property-presence profiling. For each selected entity, we first evaluate the availability of a predefined set of candidate properties spanning biographical, familial, educational, professional, and sociopolitical information. The full list of properties is available in the code source (33 in total).

We encode property availability as a binary matrix indicating whether at least one statement for the property exists in Wikidata for each entity. We exclude the properties languages spoken, written, or signed (`wdt:P1412`), writing language (`wdt:P6886`) to avoid any biased inferences during evaluation.

Popularity filtering. In step 1, we retain only entities with at least 9 Wikipedia sitelinks. This ensures a minimum level of notability and multilingual presence, and increases the likelihood that the entities and their associated property values are labeled across the selected language set.

Property selection and coverage scoring. To select the final property set, we formulate a set cover optimization problem. Let L be the set of target languages and let S denote a candidate property set. For each language $\ell \in L$, we compute $\text{Coverage}_\ell(S)$, i.e., the number of entities in language subset ℓ for which all properties in S are available. We then rank candidates according to:

$$\text{Score}(S) = |S| \cdot \sum_{\ell \in L} \text{Coverage}_\ell(S) \qquad (1)$$

[1] Dataset and Code for this paper are available at https://github.com/duo0301/WILA-popQA

[2] wikidata-truthy-28.10.23.hdt. The Pre-generated HDT files for Wikidata we used is taken from https://qanswer-svc4.univ-st-etienne.fr/

Language	Curated set of variants (Wikidata items)
English	wd:Q1860, wd:Q7976, wd:Q7979, wd:Q44676, wd:Q44679, wd:Q7053766, wd:Q48767245
Arabic	wd:Q13955, wd:Q29919, wd:Q56499, wd:Q1194795, wd:Q1654327, wd:Q5329979
German	wd:Q188, wd:Q248682, wd:Q306626, wd:Q106937689, wd:Q26721, wd:Q387066
French	wd:Q150, wd:Q1450506, wd:Q214086, wd:Q3083193, wd:Q979914, wd:Q83503
Italian	wd:Q652
Polish	wd:Q809
Hindi	wd:Q1568
Russian	wd:Q7737, wd:Q608923
Chinese	wd:Q7850, wd:Q24841726, wd:Q13414913, wd:Q18130932, wd:Q100148307

Table 1: Mapping from each target language to a curated set of Wikidata language variants, represented as items (QIDs), used to build language-specific subsets via the wdt:P1412 property.

where $|S|$ is the number of properties in the set. We also report the average coverage:

$$\text{AvgCoverage}(S) = \frac{1}{|L|} \sum_{\ell \in L} \text{Coverage}_\ell(S) \quad (2)$$

Since $|L|$ is fixed in our setting, ranking by Eq. 1 is equivalent to ranking by $|S| \cdot \text{AvgCoverage}(S)$. This criterion favors property sets that remain large while preserving coverage for as many entities as possible across languages.

Although the highest-scoring configuration corresponds to a 4-property set (*S1* in Table 2), we ultimately select the 6-property set *S3*, consisting of date of birth (wdt:P569), place of birth (wdt:P19), country of citizenship (wdt:P27), occupation (wdt:P106), date of death (wdt:P570), and place of death (wdt:P20). The inclusion of date and place of death restricts the evaluation to deceased entities; the ethical considerations underlying this decision are detailed in Section 6.1.

We report in Table 3 the number of entities per language that satisfy the data collection constraints and their coverage with respect to selected Wikidata properties. For each property, the corresponding column reports the number of entities for which the property is present and satisfies the label-coverage condition we defined before.

Furthermore, to ensure unbiased and fair evaluation, we apply additional steps to our dataset.

Entity and property completeness. After step 3, we assess completeness w.r.t the target language set and the selected property set. A property is complete w.r.t an entity if at least one of its values has labels in all target languages. An entity is considered *complete* if this condition holds for every selected property. We report the number of complete entities in Table 3.

Popularity distribution matching. We apply an additional balancing step to align entity popularity across language subsets. This step is applied to the complete entities in our dataset to mitigate popularity-related confounding effects in multilingual LLM evaluation.

Entity popularity correlates with both data completeness and model performance. Therefore, differences in sitelink distributions (strong imbalances observed in the initial curated data with the complete entities, Figure 1a) across languages can bias cross-lingual comparisons. To address this, we apply a *hard distribution-matching* procedure based on sitelink counts.

We discretize sitelink counts into fixed-width bins of size 5, with the same bin intervals across all languages. For each bin, we compute the minimum number of entities across languages and sample it from each language subset in that bin. This produces balanced evaluation subsets with sitelinks distributed equally (cf. Figure 1b).

We use fixed-width bins to preserve the absolute scale of popularity and to match entities with comparable cross-lingual counts. Thus, Table 3 reports corpus-level statistics before popularity matching. After matching, each language subset contains 231 entities (2,079 in total), which form the final evaluation set.

3.2. Multilingual Large Language Model

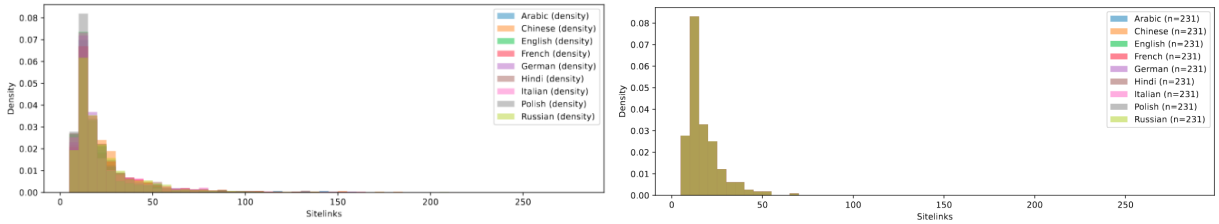
The selection of multilingual large language models for evaluation was driven by openness and reproducibility criteria. We evaluated a set of widely recognized open-weight LLMs with 7B to 16B total parameters, including 10 dense models and 2 models with mixture-of-experts (MoE) architectures. The full list of models covers *OLMo-3-7B-Instruct* (Olmo, 2025), *Mistral-7B-Instruct-v0.3* (Albert et al., 2023), *Meta-Llama-3.1-8B-Instruct* (Meta, 2024), *Qwen3-8B*, *Qwen3-14B* (Qwen, 2025), *Gemma-2-9b-it* (Gemma, 2024), *NVIDIA-Nemotron-Nano-9B-v2* (NVIDIA, 2025), *Glm-4-9b-chat-hf* (GLM, 2024), *Gemma-3-12b-it* (Gemma, 2025), *Phi-4* (Microsoft.Research, 2024), *DeepSeek-V2-Lite-Chat* (DeepSeek-AI, 2024), *Moonlight-16B-A3B-Instruct* (Moonshot-AI, 2025). For concision in the heatmap, we denote MoE models by their total parameter count, e.g., Moonlight-16B rather than Moonlight-16B-A3B.

ID	Property Set	Size	PL	RU	AR	HI	ZH	IT	FR	EN	DE	Average	Score
S1	Occ, CoC, DoB, PoB	4	12051	20173	7158	5646	7695	31531	58573	128514	79054	38932.78	1401580
S2	Occ, CoC, DoB, Pos	4	4290	5892	2370	6392	13395	16995	83866	74660	39055	27435.00	987660
S3	**DoD, PoD, PoB, DoB, Occ, CoC**	**6**	**5136**	**9357**	**2984**	**1307**	**1569**	**14717**	**37078**	**52266**	**39285**	**18188.78**	**982194**
S4	DoD, DoB, PoB, Occ, CoC	5	6059	10353	3637	1969	2790	16666	39654	65229	49295	21739.11	978260
S5	Occ, CoC, DoD, DoB	4	6687	11171	4272	3405	10263	17067	41979	86407	51715	25885.11	931864

Table 2: Top-5 candidate property sets ranked by the coverage-based score. The selected configuration used in this study is highlighted in bold (S3).

Language	#Entities	#Complete	PoD (P20)	PoB (P19)	Occup. (P106)	CoC (P27)
English	14376	6075	9746	8581	14376	14329
French	4747	1503	2756	2390	4746	4577
German	4577	1549	2928	2395	4577	4257
Russian	2242	1106	1891	1297	2242	2220
Italian	1640	689	1284	1068	1640	1249
Arabic	853	514	720	607	853	816
Polish	847	345	647	444	847	755
Chinese	632	267	509	326	632	618
Hindi	501	275	462	296	501	501

Table 3: WILA-PopQA Statistics. **#Entities** is the number of entities per language (minimum 9 Wikipedia sitelinks). We have property coverage for the selected property set. All collected entities have a date of birth (`wdt:P569`) and a date of death (`wdt:P570`), so we do not report these as separate columns.



(a) sitelinks density of all complete entities.

(b) hard-matched sitelinks density of sampled entities.

Figure 1: Sitelinks distributions before and after hard sitelinks distribution matching in WILA-PopQA

4. Proposed approach

In this study, we propose an approach to evaluate cultural bias in multilingual LLMs by varying the alignment between the prompt language and the entity's cultural reference language. For that, we established two scenarios:

Unaligned Prompt and Entity Language. Here, Tuples were created by pairing a prompt in one language with a cultural entity in another. This configuration tests whether prompt language influences the accuracy of culturally-grounded factual retrieval, e.g., a question written in Arabic paired with an English-speaking scientist.

Aligned Prompt and Entity Language. In this scenario, tuples were constructed by pairing a prompt written in language X with a cultural entity whose native or reference language is also X. For instance, an Italian-language prompt querying information about an Italian-speaking writer. This configuration serves as a reference condition against which the unaligned scenario can be compared, isolating the effect of prompt-language alignment on factual information retrieval accuracy.

Prompt Template. The template for the multilingual prompts is uniform across all languages to eliminate variations in wording, structure, or information order, ensuring a fair model comparison. The prompt was originally written in English and translated into the remaining 8 languages by native speakers, adapting phrasing to suit each language rather than using a word-for-word translation. The prompts used for all 4 properties across all languages are provided [3].

Each prompt begins with an instruction part followed by the question for the entity-property pair. The instructions vary slightly across properties to control the format of the generated answers, facilitating easier comparison with the Wikidata ground truth. For example, the templates for place of birth and date of birth include the instructions *Return the city name as the answer* and *Return answers in 'YYYY-MM-DD' format*, respectively, because LLMs' initial responses were inconsistent. Similarly, *Answer must follow this format: [answer1, answer2,...]* was included for occupation and country of citizenship to accommodate more than one answer. The country of citizenship also includes

[3] via https://zenodo.org/records/19249706 in `Prompts/prompt_template`

an example part to distinguish between the correct and incorrect format. Naturally, the question part of the prompt is tailored to each property, with a placeholder for the entity label in the prompt language, or in English if the label is absent from Wikidata for that language.

Large Language Model Experiments All models were evaluated under a fixed configuration without hyperparameter tuning, as the objective is to assess out-of-the-box factual retrieval performance across languages rather than to optimize model-specific accuracy. Each property was queried in a separate prompt to avoid cross-property interference in the model's output. Where a model exposes a system role, the prompt instruction was placed there, and the property question was submitted via the user role; for models without a system role, instruction and question were concatenated into a single input. For each model, a single inference run was performed for each (entity, property, language of the question) tuple.

Property-Specific Evaluation. Given the heterogeneous nature of the properties under evaluation, each required a tailored assessment pipeline.

Occupation. Ground-truth occupation labels are available across all target languages, but equivalent occupations may differ substantially in surface form (e.g., synonymy, inflection, or multi-word variants). To reduce reliance on exact string matching, we score occupation answers with BLEURT (Sellam et al., 2020), a learned semantic similarity metric that yields a graded similarity signal between the model output and the reference label(s). BLEURT scores natively range from −1 to 1; we rescale them to [0, 1] to enable a shared scale across all four properties in the reported figures.

Country of Citizenship. Model outputs are first normalized from any language to English using the Wikidata API[4], with DeepSeek-V3 671B as a fallback, resolving demonyms to canonical country names. Matching then proceeds through an ordered cascade of six levels: exact string match (score 1.0), alias match against Wikidata `skos:altLabel` and `rdfs:label` (0.9), bidirectional substring match for strings of four or more characters (0.85), demonym match against property P1549 (0.8), historical match resolving predecessor states via Wikidata `P1366 successor` and `P17 country` chains (0.75), and no match (0.0). Cases that fail all string-based levels are submitted to DeepSeek-V3 671B, chosen as the reference LLM judge, which receives the original question, ground truth, known aliases, and demonyms to produce a final classification.

Place of Birth. Model outputs are normalized following the same procedure as P27. Matching proceeds in two passes. In the first pass, outputs are compared against the ground truth by case-folded exact or substring matching (score 1.0); where string matching fails, Wikidata `wbsearchentities` resolves place names to QIDs, with QID equality treated as an exact match (1.0). Outputs sharing the same country as the ground truth via a SPARQL P17 query receive a partial score (0.8). Remaining unresolved cases are submitted in a second parallel pass to DeepSeek-V3 671B, which classifies them into: exact match (same place, different transliteration or language, score 1.0), country match (same country, different city, 0.8), historical match (e.g. Leningrad → St. Petersburg, 0.7), LLM-confirmed match (0.65), or no match (0.0).

Date of Birth. Dates are normalized to ISO 8601 format (YYYY-MM-DD) and evaluated through a rule-based string matching procedure without external API calls. Matching is scored as follows: a full match on year, month, and day receives a weight of 1.0; a match on year and month without day agreement receives a weight of 0.7; and a year-only match receives a weight of 0.5.

Evaluation of LLM-as-a-Judge. To validate the DeepSeek-V3 judge used as a fallback in the `Country of Citizenship` and `Place of Birth` evaluation pipelines, we conducted a human annotation study on 198 sampled judge decisions, double-annotated across all nine languages. Results are reported in Section 5.

5. Evaluation and Results

This section reports the main findings of the multilingual evaluation across four factual properties [5]. We analyze performance as a function of the **Language of Question** (LoQ), i.e., the language used to formulate the prompt, and the **Language of Entity** (LoE), i.e., the language spoken, written, or signed by the entity. Results are reported at three levels: **LoE** × **LoQ** weighted interactions (averaged across models), **Model** × **LoE**, and **Model** × **LoQ**.

The evaluation compares LLM responses across LoQ and LoE conditions, using controlled entity sampling and property-specific accuracy metrics.

Figure 2 reports the LoE × LoQ weighted interactions averaged across all models for `date of birth`, `place of birth`, `country of citizenship`, and `occupation`. Column-wise variance generally exceeds row-wise variance, indicating that LoQ is a strong performance determinant, though the relative contribution of LoQ and LoE varies by property (Table 12). The diagonal cells, representing matched LoQ–LoE conditions, do not correspond to the row maximum for any LoE group across the three properties, indicating that

[4] https://www.mediawiki.org/wiki/Wikibase/API

[5] Under `Experiments` and `Evaluation` folders

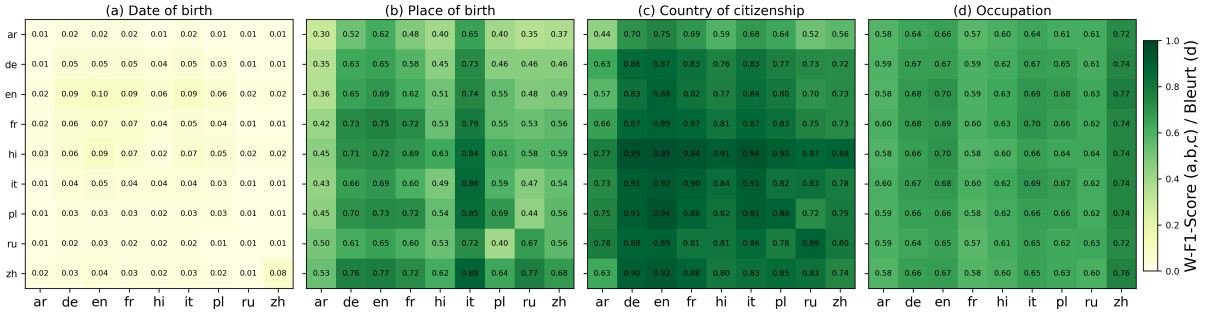


Figure 2: Heatmap for LoE (Rows) × LoQ (columns) configurations. Scores averaged over all models. Scores are weighted F1 for properties (a–c) and BLEURT rescaled to [0, 1] for (d)

Figure 3: Heatmap for Models (Rows) × LoE (columns) configurations. Scores averaged over all LoQ. Scores are weighted F1 for properties (a–c) and BLEURT rescaled to [0, 1] for (d)

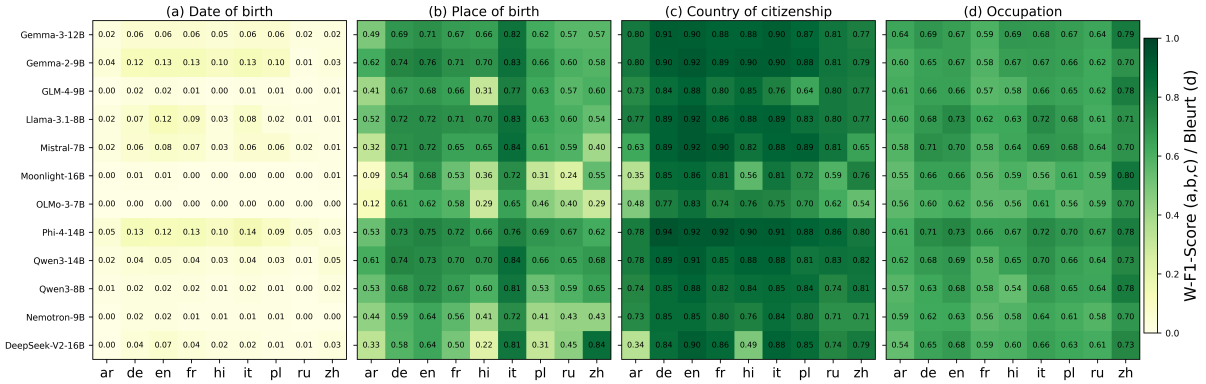


Figure 4: Heatmap for Models (Rows) × LoQ (columns) configurations. Scores averaged over all LoE. Scores are weighted F1 for properties (a–c) and BLEURT rescaled to [0, 1] for (d)

prompting a model in the entity's language does not reliably improve factual retrieval accuracy. The ar column is consistently the weakest LoQ condition across all rows and properties. The property-specific LoQ pivot varies: English and German are the strongest query languages for `country of citizenship`; Italian dominates for `place of birth` across all entity-language rows; German and English lead for `date of birth`. The zh × zh cell for `date of birth` is notably elevated relative to other cells in the *zh* row (discussed under RQ3), but does not alter the general pattern. Results for `occupation`, evaluated via BLEURT, are reported in Figure 2.

Figure 3 reports weighted F1 per model as a function of LoE. The LoE effect is secondary to LoQ but consistent: Arabic entities produce the lowest scores across all models and properties, while Hindi entities yield the highest scores for `country of citizenship` and Chinese entities lead for `place of birth`. Models with reported Chinese-dominant training, namely Qwen3-14B and Qwen3-8B, exhibit a stronger advantage for Chinese-language entities.

Model	DoB (F1)	PoB (F1)	Country (F1)	Occupation (BLEURT)	Avg. Var (No DoB)	Avg. Var (w/ DoB)
Gemma-3-12B	0.046 ± 0.018	0.644 ± 0.090	0.858 ± 0.048	0.344 ± 0.100	0.00684	0.00521
Gemma-2-9B	0.088 ± 0.045	0.689 ± 0.077	0.866 ± 0.048	0.294 ± 0.068	<u>0.00426</u>	0.00370
GLM-4-9B	0.009 ± 0.007	0.589 ± 0.135	0.786 ± 0.068	0.284 ± 0.112	0.01180	0.00887
Llama-3.1-8B	0.050 ± 0.038	0.663 ± 0.093	0.846 ± 0.052	0.332 ± 0.092	0.00665	0.00536
Mistral-7B	0.046 ± 0.024	0.610 ± 0.152	0.821 ± 0.103	0.313 ± 0.094	0.01411	0.01073
Moonlight-16B	0.003 ± 0.005	0.447 ± 0.198	0.701 ± 0.160	0.241 ± 0.150	0.02902	0.02177
OLMo-3-7B	0.000 ± 0.000	0.447 ± 0.174	0.688 ± 0.110	0.193 ± 0.090	0.01685	0.01263
Phi-4	**0.093 ± 0.039**	0.681 ± 0.068	**0.879 ± 0.053**	**0.389 ± 0.091**	0.00526	0.00431
Qwen3-14B	0.034 ± 0.013	**0.701 ± 0.062**	0.850 ± 0.040	0.320 ± 0.088	0.00439	<u>0.00333</u>
Qwen3-8B	0.014 ± 0.007	0.642 ± 0.086	0.819 ± 0.046	0.277 ± 0.136	0.00929	0.00698
Nemotron-9B	0.009 ± 0.007	0.503 ± 0.109	0.783 ± 0.050	0.224 ± 0.081	0.00686	0.00517
DeepSeek-V2-16B	0.027 ± 0.020	0.520 ± 0.193	0.743 ± 0.166	0.263 ± 0.116	0.02906	0.02186

Table 4: Performance and robustness across tasks. Values are mean ± standard deviation across the 9 Languages of Question (LoQ: ar–zh).DoB, PoB, and Country are weighted F1 scores; Occupation is BLEURT. The last two columns report the average variance across tasks, computed with and without DoB. **Bold** indicates best performance per task; <u>underline</u> indicates best stability.

Figure 4 reports *weighted F1* per model as a function of *LoQ*. Arabic as *LoQ* is the weakest condition across all three properties, with the most pronounced reductions observed in *Moonlight-16B* and *OLMo-3-7B*. The ranking of remaining query languages is property-dependent, consistent with the LoE × LoQ analysis above. Table 4 reports per-model performance and cross-lingual stability. *Phi-4* achieves the highest scores on three of four properties, while *Qwen3-14B* exhibits the greatest overall cross-lingual stability (lowest average variance across *LoQ*). Because *DoB* shows near-zero variance across languages, we also report stability results excluding *DoB*, in which *Gemma-2-9B* becomes the most stable model. *Moonlight-16B* and *DeepSeek-V2-16B* exhibit the highest dispersion in both settings, consistent with their lower overall performance.

Finally, we validate the DeepSeek-V3 judge fallback through a human annotation study on 198 stratified samples across all 9 languages and both verdict types. The judge achieves 91.4% agreement with primary annotators and 92.9% with secondary annotators, with substantial inter-annotator agreement (Cohen's κ = 0.755). The few disagreements primarily concern historical state transitions (see Tables 15 and 16 in the Appendix).

6. Discussion and Conclusions

This study proposed a Wikidata-based framework for probing multilingual LLMs on culturally grounded factual properties, based on the language of the question (LoQ) and the language of the entity (LoE), and measuring entity popularity within a single evaluation pipeline. Three research questions were addressed in turn.

RQ1: How does LoQ affect factual accuracy in multilingual LLMs for entities tied to specific linguistic and cultural contexts? The results confirm that LoQ is a major determinant of factual retrieval accuracy, and the dominant factor for `place of birth` (η^2 = 0.627) and `occupation` (η^2 = 0.911). For `country of citizenship`, LoQ and LoE contribute comparably(Table 12). Column-wise variance in the LoE ×LoQ interaction matrices generally exceeds row-wise variance, though the relative contribution varies by property. The direction of the LoQ effect is property-dependent: English and German constitute the strongest query languages for `country of citizenship`; Italian is the strongest query language for `place of birth` across all entity-language rows and all models without exception; German and English lead for `date of birth`, although its scores remain near-zero across all LoQ conditions. Arabic as LoQ is the weakest condition across all properties and all models, with the most pronounced score reductions observed for Moonlight-16B and OLMo-3-7B. Importantly, the hypothesis that aligning the query language with the entity's language (i.e., the matched LoQ–LoE diagonal) yields higher factual retrieval accuracy is not supported by the data: diagonal cells do not correspond to row-maxima for any property, and in several cases, most markedly for Arabic-language entities, querying in English or German produces substantially higher scores than querying in the entity's language.

RQ2: Can a Wikidata-centric methodology measure such language effects while controlling for confounds related to LoE and data availability? The sitelink-equidistributed sampling strategy enables the analysis of LoQ and LoE effects in relative isolation from entity-popularity artifacts. The consistency of these effects across 12 models, 4 properties, and heterogeneous evaluation pipelines (string matching, API normalization, BLEURT) supports the internal validity of the methodology. The LoE effect, while secondary to LoQ, is measurable and

[5]The η^2 gap between LoE and LoQ is narrower than for other properties; thus, the two factors contribute comparably.

directionally consistent: Arabic-language entities yield the lowest weighted F1 scores across models and properties, while Hindi entities perform comparatively higher for `country of citizenship` and Chinese entities for `place of birth`. The reproducibility of these patterns across distinct evaluation pipelines indicates that the observed differences are not artifacts of a single metric or matching procedure.

Coverage limitations due to missing multilingual labels in Wikidata were explicitly documented at each evaluation step. In cases where a property value lacked a label in the query language, the evaluation either documented the absence or fell back to a documented proxy (an English label or LLM-assisted normalization), preserving the traceability of each decision. This transparency supports interpretable error analysis and enables replication under alternative coverage assumptions.

RQ3: Which quantifiable factors help explain performance differences after controlling for LoQ and LoE? Here, sitelink popularity served as the primary covariate to normalize for entity selection effects. The consistency of LoQ effects across equidistributed entity-popularity groups indicates that the observed language-level differences are not attributable to differential entity prominence across language subsets, addressing a known confound in prior multilingual probing work (Wiland et al., 2024). The residual LoE effect — most pronounced for Arabic-language entities and for models trained on Chinese-dominant data (Qwen3-14B, Qwen3-8B), suggests that the composition of the training data is an additional explanatory factor. However, the present design does not permit this to be quantified directly, as model-specific corpus statistics are not publicly available for all evaluated systems.

The property type we prompt the model for constitutes an independent explanatory factor. The consistently low DoB scores reflect a limitation in factual recall. Across all evaluations (276,480 samples), 88.1% of outputs conform to the required format, yet 89.7% of those contain incorrect dates (see Table 13 in Appendix). Partial matches follow a granularity gradient, with year-only matches (7.2%) exceeding year-and-month (1.2%) and exact matches (1.0%), suggesting that LM operates as a "close-enough" semantic retriever and fails on precise numeric values. This is consistent with findings that LMs encode numeric properties, such as birth year, along continuous, monotonic directions in activation space, capturing approximate temporal regions without pinpoint precision (Heinzerling and Inui, 2024). More broadly, Wikidata properties differ in how LMs recall them. Some are well-memorized due to their frequency in the training data, while others, such as nationality, can be inferred from surface-level artifacts in entity names (Mallen et al., 2023). These explain the comparatively higher scores observed for the country of citizenship in our evaluation. Separately, the *zh* ×××× *zh* cell for `date of birth` (Fig. 2) is notably elevated relative to other cells in the zh row, likely driven by the training data composition of the Qwen3 models.

General conclusions Across all research questions, the evaluation produces three findings. First, LoQ is a dominant factor in multilingual factual question answering over culturally grounded entities; its relative importance versus LoE is property-dependent, and the direction of the LoQ effect does not uniformly favor English or the entity's language. Second, aligning the query language with the entity's language does not reliably improve factual retrieval accuracy and, in low-resource language conditions, can substantially reduce it. Third, sitelink popularity normalization does not fully eliminate cross-language performance differences, indicating that the composition of the training data constitutes an additional explanatory variable beyond entity selection effects.

These findings have direct implications for KG+LLM workflows that rely on multilingual factual retrieval: the choice of query language should be treated as an explicit design parameter, with property-specific evidence considered when selecting the prompt language, rather than defaulting to either English or the entity's language.

Future work should address three limitations of the present study. First, extending the entity set to include languages with lower Wikidata coverage, particularly those lacking multilingual labels for a substantial proportion of property values, would test the robustness of the observed LoQ effects under more severe data sparsity conditions. This extension should also include widely spoken languages not covered by the current author team, notably Spanish, which spans both European and Latin American cultural contexts and would enable cross-cultural comparisons within a single language. Second, incorporating direct measures of training data composition (when available) would allow the residual LoE effect to be decomposed into contributions from label coverage and pretraining corpus. Third, applying the framework to generative tasks beyond factual slot-filling, such as open-ended biographical description or relation extraction, would determine whether the observed LoQ patterns extend to settings where ground truth is less constrained.

6.1. Ethical considerations and limitations

This work uses publicly available Wikidata statements about human entities. To reduce legal and ethical risks associated with processing personal data of living persons under EU data-protection law (GDPR Recital 27) [6], we restrict the entity set to deceased individuals. We implement this by requiring values for date of death (`wdt:P570`) and place of death (`wdt:P20`). We then accept the resulting coverage trade-off in Table 2 and select *S3* rather than *S1* as our property set. This restriction additionally reduces potential issues arising from temporal lag between living individuals acquiring new occupations or languages and the subsequent reflection of these changes in web resources and Wikidata.

We also exclude sex or gender (`wdt:P21`) from the retrieved properties. This prevents the inclusion of sensitive personal attributes and it acknowledges that `wdt:P21` does not authentically reflect the complexity of human gender identities, limiting its suitability as ground truth (Melis et al., 2025). More generally, restricting the set of relations/properties is common in knowledge-base benchmarks (e.g., SimpleQuestions (Bordes et al., 2015)) and data-to-text corpora (e.g., WebNLG (Gardent et al., 2017)).

Limitations. Restricting the sample to deceased individuals biases it toward representing well-documented historical figures and underrepresenting contemporary figures. This can limit the extent to which results generalize to QA about current questions.

7. Bibliographical References

References

Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. 2023. Mistral 7B.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models.

Hiba Arnaout, Trung-Kien Tran, Daria Stepanova, Mohamed Hassan Gad-Elrab, Simon Razniewski, and Gerhard Weikum. 2022. Utilizing language model probes for knowledge graph repair. In *Wiki Workshop 2022*.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *ArXiv*, abs/1506.02075.

Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphael Romero, Iman Johary, Alexandru-Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. 2024. Large Language Models Reflect the Ideology of their Creators.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *CoRR*, abs/2405.04434.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Team Gemma. 2024. Gemma 2: Improving open language models at a practical size.

Team Gemma. 2025. Gemma 3 technical report.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

Ming Jiang and Mansi Joshi. 2024. CPopQA: Ranking Cultural Concept Popularity by LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 615–630, Mexico City, Mexico. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for*

[6] https://gdpr-info.eu/recitals/no-27/

Computational Linguistics: Main Volume, pages 3250–3258, Online. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models.

David Kubeša and Milan Straka. 2023. Damuel: A large multilingual dataset for entity linking.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Jan Malte Lichtenberg, Alexander Buchholz, and Pola Schwöbel. 2024. Large Language Models as Recommender Systems: A Study of Popularity Bias. ArXiv:2406.01285 [cs].

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

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, Toronto, Canada. Association for Computational Linguistics.

Beatrice Melis, Marta Fioravanti, Chiara Paolini, and Daniele Metilli. 2025. How have you modelled my gender? reconstructing the history of gender representation in wikidata. *Internet Histories*, 9(1-2):163–179.

Meta. 2024. The Llama 3 Herd of Models.

Microsoft.Research. 2024. Phi-4 Technical Report.

Moonshot-AI. 2025. Muon is scalable for llm training.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A Dataset for Cross-Lingual Comparison of Stereotypes in Generative LLMs.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2025. How Knowledge Popularity Influences and Enhances LLM Knowledge Boundary Perception. ArXiv:2505.17537 [cs].

NVIDIA. 2025. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model.

Team Olmo. 2025. Olmo 3.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pages 2463–2473.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023a. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023b. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models.

Team Qwen. 2025. Qwen3 technical report.

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications.

John Samuel. 2021. Wdprop: Web application to analyse multilingual aspects of wikidata properties. In *Proceedings of the 17th International Symposium on Open Collaboration*, OpenSym '21, New York, NY, USA. Association for Computing Machinery.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

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, Mexico City, Mexico. Association for Computational Linguistics.

Kim Trong Vu, Michael Krumdick, Varshini Reddy, Franck Dernoncourt, and Viet Dac Lai. 2024. An analysis of multilingual factscore.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual prompting for improving LLM generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6367–6389, Suzhou, China. Association for Computational Linguistics.

Wikidata contributors. 2026. Wikidata: Database Reports/List of Properties/All.

Jacek Wiland, Max Ploner, and Alan Akbik. 2024. BEAR: A unified framework for evaluating relational knowledge in causal and masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2393–2411, Mexico City, Mexico. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models?

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605, Singapore.

A. Appendix

A.1. Sitelinks hard matching targets

Sitelinks bin	Target count
[5–10)	32
[10–15)	96
[15–20)	38
[20–25)	29
[25–30)	14
[30–35)	7
[35–40)	7
[40–45)	3
[45–50)	2
[50–55)	2
[65–70)	1

Table 5: Hard matching targets per fixed-width sitelinks bin. The target count equals the minimum number of available complete entities across languages each bin; we sample this many entities per language.

A.2. Most Represented Occupations

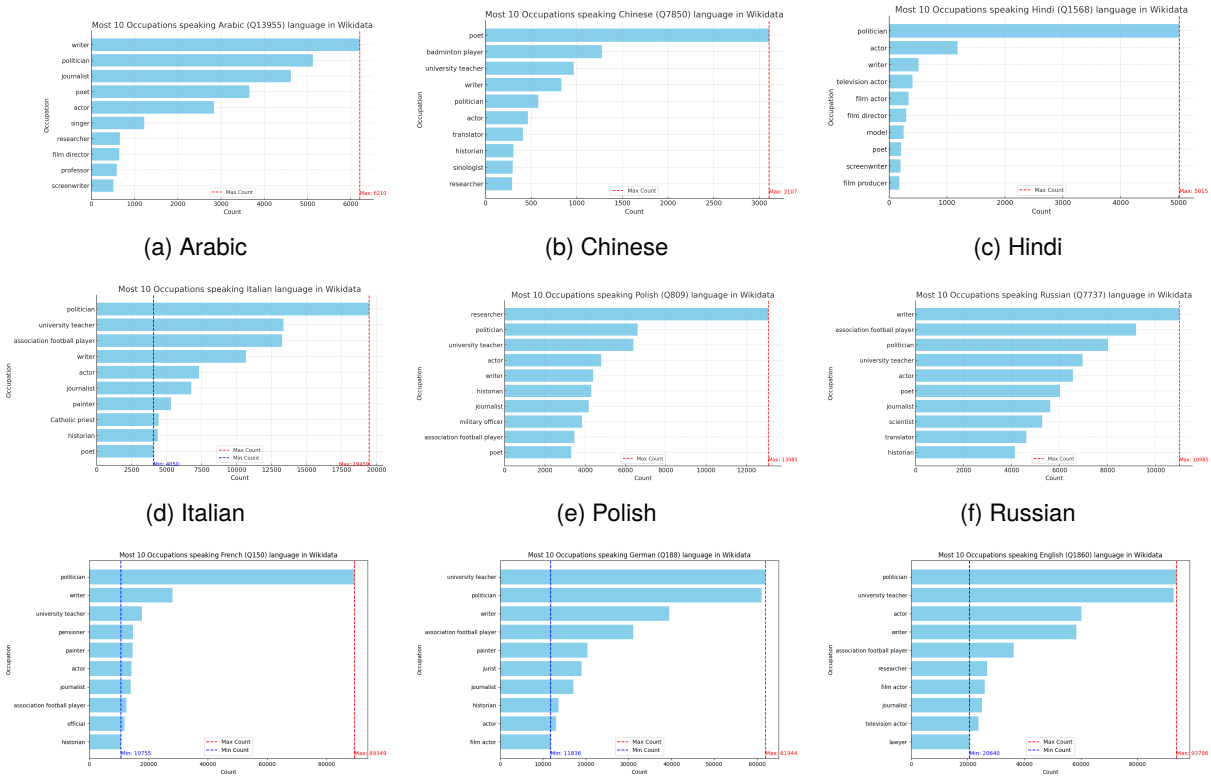


Figure 5: Top-10 occupations for each language.

Across the language subsets for which we currently report top-10 occupations, the intersection of the top-10 lists is {*writer*, *politician*, *actor*, *poet*}. Our occupation filter uses *writer* (`wd:Q36180`), *politician* (`wd:Q82955`), *actor* (`wd:Q33999`), and the broader class *creator* (`wd:Q2500638`). The difference is therefore *creator* vs. *poet*. This is consistent with Wikidata's class hierarchy: *poet* (`wd:Q49757`) is a subclass of *writer* (`wd:Q36180`); *writer* is a subclass of *author* (`wd:Q482980`); and *author* is a subclass of *creator* (`wd:Q2500638`). We therefore include *creator* as a general class that also covers *poet* through this chain.

Rank	Arabic	Chinese	Hindi	Russian	Polish
1	***writer***	poet	***politician***	***writer***	*researcher*
2	***politician***	*badminton player*	***actor***	*association football player*	***politician***
3	*journalist*	*university teacher*	***writer***	***politician***	*university teacher*
4	poet	***writer***	*television actor*	*university teacher*	***actor***
5	***actor***	***politician***	*film actor*	***actor***	***writer***
6	*singer*	***actor***	*film director*	poet	*historian*
7	*researcher*	*translator*	*model*	*journalist*	*journalist*
8	*film director*	*historian*	poet	*scientist*	*military officer*
9	*professor*	*sinologist*	*screenwriter*	*translator*	*association football player*
10	*screenwriter*	*researcher*	*film producer*	*historian*	poet

Table 6: Top-10 occupations for **Arabic**, **Chinese**, **Hindi**, **Russian**, and **Polish**, ordered by descending frequency. The occupations intersection used in our study (*writer*, *politician*, *actor*) are highlighted in bold.

Rank	Italian	French	English	German
1	***politician***	***politician***	***politician***	*university teacher*
2	*university teacher*	***writer***	*university teacher*	***politician***
3	*association football player*	*university teacher*	***actor***	***writer***
4	***writer***	*pensioner*	***writer***	*association football player*
5	***actor***	*painter*	*association football player*	*painter*
6	*journalist*	***actor***	*researcher*	*jurist*
7	*painter*	*journalist*	*film actor*	*journalist*
8	*catholic priest*	*association football player*	*journalist*	*historian*
9	*historian*	*official*	*television actor*	***actor***
10	poet	*historian*	*lawyer*	*film actor*

Table 7: Top-10 occupations for **Italian**, **French**, **English**, and **German**, ordered by descending frequency. The occupations intersection used in our study (*writer*, *politician*, *actor*) are highlighted in bold.

A.3. Model sources

Company	Model	Parameters	Arch.	Release	K.cutoff
DeepSeek	*DeepSeek-V2-Lite-Chat* (DeepSeek-AI, 2024)	16B, 2.4B (a)	MoE	May 2024	–
Moonshot AI	*Moonlight-16B-A3B-Instruct* (Moonshot-AI, 2025)	16B, 3B (a)	MoE	Feb 2025	–
Allen AI	*OLMo-3-7B-Instruct* (Olmo, 2025)	7B	Dense	Dec 2025	Dec 2024
Mistral AI	*Mistral-7B-Instruct-v0.3* (Albert et al., 2023)	7B	Dense	May 2024	–
Meta	*Meta-Llama-3.1-8B-Instruct* (Meta, 2024)	8B	Dense	Jul 2024	Dec 2023
Alibaba	*Qwen3-8B* (Qwen, 2025)	8B	Dense	May 2025	–
Google	*Gemma-2-9b-it* (Gemma, 2024)	9B	Dense	Jun 2024	–
NVIDIA	*NVIDIA-Nemotron-Nano-9B-v2* (NVIDIA, 2025)	9B	Dense	Aug 2025	Sep 2024
Zhipu AI	*Glm-4-9b-chat-hf* (GLM, 2024)	9B	Dense	Jun 2024	–
Google	*Gemma-3-12b-it* (Gemma, 2025)	12B	Dense	Mar 2025	Aug 2024
Alibaba	*Qwen3-14B* (Qwen, 2025)	14B	Dense	May 2025	–
Microsoft	*Phi-4* (Microsoft.Research, 2024)	15B	Dense	Dec 2024	Jun 2024

Table 8: Language models used in our experiments. K.cutoff denotes the (reported) training knowledge cutoff when available; "–" denotes information not disclosed by the provider. Entries marked with (a) indicate the number of activated parameters in the MoE models architecture (Arch.) during inference.

B. Wikidata Candidate Properties

Category	Wikidata Properties
Biographical	wdt:P19 (place_of_birth) wdt:P20 (place_of_death) wdt:P569 (date_of_birth) wdt:P570 (date_of_death) wdt:P509 (cause_of_death) wdt:P119 (place_of_burial) wdt:P551 (residence)
Family / Relationships	wdt:P22 (father) wdt:P25 (mother) wdt:P26 (spouse) wdt:P40 (child) wdt:P3373 (sibling) wdt:P1066 (student_of) wdt:P184 (doctoral_advisor)
Career / Professional	wdt:P39 (position_held) wdt:P101 (field_of_work) wdt:P106 (occupation) wdt:P108 (employer) wdt:P135 (movement) wdt:P166 (award_received) wdt:P463 (member_of) wdt:P512 (academic_degree) wdt:P800 (notable_work) wdt:P937 (work_location)
Actor-specific	wdt:P161 (cast_member_of) wdt:P453 (character_role) wdt:P725 (voice_actor)
Sociocultural / Misc.	wdt:P27 (country_of_citizenship) wdt:P69 (educated_at) wdt:P103 (native_language) wdt:P140 (religion) wdt:P172 (ethnic_group) wdt:P102 (member_of_political_party)

Table 9: Grouped candidate Wikidata properties considered during data collection.

C. Prompt templates

C.1. Property-specific prompts

Property	Prompt Template
place_of_birth (P19)	**##Instruction** Answer the question in English. Return only the answer keyword. Do not include any explanations. Return the city name as answer. **##Question** What is the birthplace of **<entity_name>**?
date_of_birth (P569)	**##Instruction** Answer the question in English. Return only the answer keyword. Do not include any explanations. Return answers in 'YYYY-MM-DD' format. **##Question** What is the birth date of **<entity_name>**?
occupation (P106)	**##Instruction** Answer the question in English. Return only the answer keyword. Answer must follow this format: [answer1, answer2,...] Do not include any explanations. **##Question** What is the occupation of **<entity_name>**?
country_of_citizenship (P27)	**##Instruction** Answer the question in English. Return only the answer keyword. Answer must follow this format: [answer1, answer2,...] Do not include any explanations. **##Example** Q: What is the country of citizenship of Marie Curie? Incorrect format: French Correct format: [France] **##Question** What is the country of citizenship of **<entity_name>**?

Table 10: English prompt template for all 4 properties with entity placeholder (**<entity_name>**).

C.2. Multilingual prompts

English (en)	**##Instruction** Answer the question in English. Return only the answer keyword. Do not include any explanations. Return the city name as an answer. **##Question** What is the birthplace of **<entity_name>**?
Arabic (ar)	**##Instruction** أجب عن السؤال باللغة العربية. قم بالإجابة بالكلمة المفتاحية فقط. لا تقم بإرفاق أي تفسيرات. أعد اسم المدينة كإجابة. **##Question** ما هو مسقط رأس **<entity_name>** ؟
Chinese (zh)	**##Instruction** 用繁體中文回答給出的問題。 只回答關鍵詞答案。 不要給出任何解釋。 返回答案必須是城市。 **##Question** **<entity_name>**的出生地是?
French (fr)	**##Instruction** Réponds à la question en français. Retourne uniquement le mot-clé de la réponse. Aucune explication ne doit être incluse. Indiquez le nom de la ville comme réponse. **##Question** Quel est le lieu de naissance de **<entity_name>** ?
German (de)	**##Instruction** Beantworte die Frage in Deutsch. Gib die Antwort nur in Form eines Schlüsselworts zurück. Füge keine Erklärungen hinzu. Gib den Namen der Stadt als Antwort zurück. **##Question** Was ist der Geburtsort von **<entity_name>**?
Hindi (hi)	**##Instruction** प्रश्न का उत्तर हिंदी में दें। केवल उत्तर शब्द लौटाएँ। कोई स्पष्टीकरण शामिल न करें। उत्तर के रूप में शहर का नाम लौटाएँ। **##Question** **<entity_name>** का जन्मस्थान क्या है?
Italian (it)	**##Instruction** Rispondi in Italiano. Rispondi soltanto con la risposta della domanda. Non includere comment o spiegazioni alla risposta. Rispondi soltanto con il paese in cui è nata la persona. **##Question** Qual è il luogo di nascita di **<entity_name>**?
Polish (pl)	**##Instruction** Odpowiedz na pytanie w języku polskim. Podaj wyłącznie nazwę miejscowości. Nie dodawaj żadnych wyjaśnień. **##Question** Gdzie urodził/urodziła się **<entity_name>**?
Russian (ru)	**##Instruction** Отвечайте на вопрос на Русском. Возвращайте только краткий ответ в виде ключевого слова. Не добавляйте никаких объяснений. Возвращайте название города в качестве ответа. **##Question** Где родился/родилась **<entity_name>**?

Table 11: Multilingual prompt for the property place_of_birth

C.3. Algorithms

Algorithm 1: BLEURT-based scoring for multi-valued occupation fields (best-match aggregation)

Input: R: list of reference strings (ground truth); C: list of candidate strings (LLM output); $\mathcal{S}$: BLEURT scorer returning scores for paired lists
Output: Scalar score $s \in \mathbb{R}$ (BLEURT best-match score), with sentinel values

if $|C| = 0$ **then**
  return -1.0 `// No LLM output`
if $|R| = 0$ **then**
  return 0.0 `// No ground truth (unexpected)`

$R' \leftarrow [\,\text{strip}(r) \;:\; r \in R \wedge r \neq \emptyset \wedge \text{strip}(r) \neq \emptyset\,]$
$C' \leftarrow [\,\text{strip}(c) \;:\; c \in C \wedge c \neq \emptyset \wedge \text{strip}(c) \neq \emptyset\,]$
if $|R'| = 0$ ***or*** $|C'| = 0$ **then**
  return 0.0 `// No valid pairs after filtering`

`all_refs` $\leftarrow []$; `all_cands` $\leftarrow []$
foreach $r \in R'$ **do**
  foreach $c \in C'$ **do**
    append r to `all_refs`
    append c to `all_cands`

`scores` $\leftarrow \mathcal{S}$(`all_refs`, `all_cands`) `// Flat list of length` $|R'| \cdot |C'|$
`best_scores` $\leftarrow []$
$n \leftarrow |C'|$
for $i \leftarrow 0$ **to** $|R'| - 1$ **do**
  `row` $\leftarrow$ `scores`$[i \cdot n : (i+1) \cdot n]$
  append max(`row`) to `best_scores`

return max(`best_scores`)

C.4. Statistical test: Two-way ANOVA

Property	η^2_{LoE}	η^2_{LoQ}	$\eta^2_{\text{Res.}}$	F_{LoE}	F_{LoQ}	Dominant
Date of Birth	0.453	0.275	0.272	13.3	8.1	LoE
Place of Birth	0.268	0.627	0.105	20.4	47.7	LoQ
Country of Cit.	0.506	0.415	0.079	50.9	41.8	LoE
Occupation	0.055	0.911	0.035	12.5	208.2	LoQ

Table 12: Two-way ANOVA (without replication) on LoE $\times$ LoQ interaction matrices. η^2 denotes the proportion of total variance explained by each factor. All effects are significant at $p < 0.001$.

C.5. Failure Analysis of Date of Birth Evaluation Results

Category	Count	%
Overall Score Distribution (N = 276,480)		
Exact match (score = 1.0)	2,641	1.0%
Year + month match (score = 0.7)	3,358	1.2%
Year-only match (score = 0.5)	19,918	7.2%
No match (score = 0.0)	250,563	90.6%
Format Compliance		
Outputs in `YYYY-MM-DD` format	243,684	88.1%
of which correct date	25,000	10.3%*
of which wrong date	218,684	89.7%*
Failure Modes (no-match cases only, N = 250,563)		
Wrong date, correct format	218,684	87.3%
Verbose response	13,562	5.4%
Other format	11,841	4.7%
Literal format echo	5,409	2.2%
Year only, no format	896	0.4%
Empty / NaN	171	0.1%
Year Error (wrong date, correct format, N = 218,625)		
0–1 years off	5,165	2.4%
2–5 years off	17,480	8.0%
6–10 years off	21,010	9.6%
11–50 years off	88,953	40.7%
51–100 years off	53,959	24.7%
101–500 years off	29,593	13.5%
500+ years off	2,462	1.1%
Mean year error: 63.6	Median year error: 37.0	

Table 13: Failure analysis of date of birth evaluations. *Percentages relative to formatted outputs.

Model	Format %	Mean Score
Qwen3-8B	99.6	0.0432
Qwen3-14B	99.4	0.0747
Gemma-3-12B	99.3	0.0682
Gemma-2-9B	98.9	0.1041
Meta-Llama-3.1-8B	93.9	0.0526
GLM-4-9B	92.1	0.0125
OLMo-3-7B	90.5	0.0032
Nemotron-9B	86.6	0.0197
Mistral-7B	85.5	0.0885
Phi-4	84.8	0.1292
Moonlight-16B	71.1	0.0109
DeepSeek-V2-16B	56.0	0.0420

Table 14: Per-model format compliance (`YYYY-MM-DD`) versus mean DoB score. High format compliance does not translate into high accuracy, confirming that low DoB scores reflect a failure of factual recall.

D. Human Annotation Study on a sample of DeepSeek-V3 judge decisions

	Primary	Secondary
Overall	91.4%	92.9%
CoC	87.9%	89.9%
PoB	94.9%	96.0%

Table 15: DeepSeek-V3 judge validation accuracy. Percentage of cases where the human annotator agrees with the judge's verdict. Primary annotators evaluated in the LoQ language; secondary annotators used English translations.

	Primary	**Secondary**
By judge verdict		
True Positive (n=90)	90.0%	94.4%
False Positive (n=108)	92.6%	91.7%
Inter-annotator agreement		
Raw agreement	96.5%	
Cohen's κ	0.755	

Table 16: Detailed annotation study results. 198 samples (99 CoC + 99 PoB) were double-annotated across 9 languages.