Measuring the Effect of Disfluency in Multilingual Knowledge Probing Benchmarks

Kirill Semenov and Rico Sennrich
University of Zurich
kirill.semenov@uzh.ch

Abstract

For multilingual factual knowledge assessment of LLMs, benchmarks such as MLAMA use template translations that do not take into account the grammatical and semantic information of the named entities inserted in the sentence. This leads to numerous instances of ungrammaticality or wrong wording of the final prompts, which complicates the interpretation of scores, especially for languages that have a rich morphological inventory. In this work, we sample 4 Slavic languages from the MLAMA dataset and compare the knowledge retrieval scores between the initial (templated) MLAMA dataset and its sentence-level translations made by Google Translate and ChatGPT. We observe a significant increase in knowledge retrieval scores, and provide a qualitative analysis for possible reasons behind it. We also make an additional analysis of 5 more languages from different families and see similar patterns. Therefore, we encourage the community to control the grammaticality of highly multilingual datasets for higher and more interpretable results, which is well approximated by whole sentence translation with neural MT or LLM systems.[1]

1 Introduction

Large Transformer-based (Vaswani et al., 2017) pre-trained language models (LMs) are known for their fact memorization (see Petroni et al., 2019; Zhong et al., 2021). For behavioral analysis of factual knowledge retrieval of LLMs, datasets that contain facts from a particular database are used. Several datasets have been created for multilingual LM evaluation (for instance Jiang et al., 2020a; Gao et al., 2024). One of the richest datasets (both in terms of language variety and number of factual instances) is MLAMA (Kassner et al., 2021), which has been widely used by other researchers and served as a basis for further versions of multilingual knowledge benchmarks such as BMLAMA (Qi et al., 2023). While being created in the "age" of encoder LMs, it successfully survived the reorientation of the field towards the decoder-only LMs, and is being used as a probe there (see Section 2 for examples).

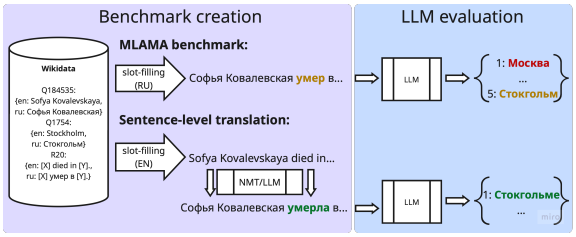

Figure 1: Main idea of our experiment. Factual dataset MLAMA uses the templated sentences in each language to prompt LLMs for factual knowledge. This results in many ungrammatical prompts: for Sofya Kovalevskaya, who was the first woman to earn a doctorate in mathematics, the verb "was born" had to be inflected in feminine gender, not in the templated masculine. This leads to poor factual extraction. Sentence-level translation leads to grammatically and lexically coherent sentences and increases factual retrieval of objects (in the correct grammatical forms, in this case, locative case of the word "Stockholm"). The incorrect entities are marked red, the correct but ungrammatical words are in yellow, the correct and grammatical words are in green.

However, if we take a closer look at the benchmarks and at the formulations of the prompts in languages other than English, we will frequently stumble upon ungrammatical sentences or wrong lexical choice. This stems from the dataset creation pipeline, which included translating the relation types and the named entities separately. Although such a procedure was clearly aimed at uniformity of the prompts, and authors report that manual template correction for German, Hindi and Japanese only had small effects on results, we revisit this question, motivated by numerous fluency issues in the data, and the shift towards decoder-only LLMs

[1]The dataset and all related code is published at the Github repository: ZurichNLP/Fluent-mLAMA.

which might show different patterns when faced with disfluent prompts. How meaningful are the resulting scores in the face of disfluency? Does a gap in scores compared to English indicate a lack of cross-lingual knowledge transfer, or do disfluent queries act as a confound that explain part of the gap?

To answer this question, we have sampled nine languages and compare factual retrieval accuracy of a Llama-2 model between existing templated sentences from MLAMA and automatic sentence-level translations. As a first set, we consider four Slavic languages which have a considerable range of grammatical categories (such as gender, case and tense markers): Russian, Czech, Ukrainian and Croatian. The languages exhibit diversity in terms of writing systems (Latin for Czech and Croatian, Cyrillics for Russian and Ukrainian), as well as their level of resourcedness. For these languages, consider two translation systems: Google NMT or (in a more controlled manner) ChatGPT with explicit named entity translation. A second set of languages aims to establish generalisation beyond Slavic languages, and includes Spanish, Chinese, Vietnamese, Indonesian and Danish. There, we compare the templated translations of the datasets with the sentence-level Google NMT translations. Our hypothesis for both experiments is that MLAMA (and other template-based benchmarks) substantially underestimate multilingual knowledge retrieval. The brief summarization of our experiment is shown in Figure 1.

Our results show significant increase in factual retrieval over most of the nine languages, especially if the model is prompted with Google NMT translation. The increase in performance, however, was not equal for all languages, and the biggest effect (up to 10%) is observed in Czech, Russian (Slavic), Vietnamese and Indonesian (non-Slavic).

With this paper, we present the main following contributions:

1. We show an empirical evidence of the claim that the grammaticality and lexical accuracy of the prompts matter for the multilingual factual evaluation of the decoder LLMs;

2. We create and publish the curated version of the MLAMA dataset of the four Slavic and five non-Slavic languages that we used for evaluation;

3. We publish the straightforward pipeline for extension of the grammatical dataset for the rest of the languages present in MLAMA.

2 Related Work

Prompt-based probing has been one of the main ways of the factual knowledge analysis of LMs. The first datasets for this task, such as zsRE (Levy et al., 2017) or ParaRel (Elazar et al., 2021), were initially created to test encoder-based models such as BERT (Devlin et al., 2018), but then started being used for decoder models evaluation (for example Yao et al., 2023; Meng et al., 2022). Older datasets often serve as a basis for new benchmarks: the LAMA dataset (Petroni et al., 2019), another popular factual dataset, was itself an updated version of the T-REx (Elsahar et al., 2018) dataset. Later, some research teams used LAMA itself as a material for filtered benchmarks like LAMA-UHN (Poerner et al., 2020), while others borrowed the principles of data collection for particular domains, such as BioLAMA (Sung et al., 2021) or Legal-LAMA (Chalkidis et al., 2023).

The datasets also demonstrate continuity in terms of multilingual factual knowledge analysis. Just as different multilingual versions of zsRE (Wang et al., 2024b) and ParaRel (Fierro and Søgaard, 2022), the LAMA dataset was translated several times, the most well-known version being MLAMA (Kassner et al., 2021). This was a significant contribution to multilingual evaluation of LMs, since the dataset was translated into 53 languages. The MLAMA benchmark has also become the source, whether in terms of material or of creation protocols, to later multilingual datasets, such as BMLAMA (Qi et al., 2023), GeoMLAMA (Yin et al., 2022) or DLAMA (Keleg and Magdy, 2023); it also widely used by the community as is until now (Zhao et al., 2024; Xu et al., 2023; Chen et al., 2023).

The challenges of prompt-based probing have been a long-standing research focus. Firstly, we cannot be sure that the prompts manually formulated by the researchers (which was the initial approach) would maximize the retrieval of a fact. Several solutions were suggested, from corpus-based retrieval of optimal prompt phrasing (Jiang et al., 2020b), to gradient-based methods, either by choosing arbitrary tokens for encoders (such as AutoPrompt, Shin et al., 2020) or by creating coherent sentences for decoders (Yin et al., 2024). Secondly, since many approaches are based on providing an

LLM with a range slot-filling (or next token) options, and ranking the correct options against the distractors, a body of work is related to finding the optimal distractor set: see (Khodja et al., 2025) for an overview.

Another set of problems is related to multilingual prompting. DLAMA and GeoMLAMA, but also xGeo and xPeo (Gao et al., 2024) are datasets that are more balanced in terms of geographical or cultural distribution of facts in the datasets. The translation protocol is another axis of variation among the multilingual datasets: the most popular way is using MT models such as Google Translate (see BizsRE (Wang et al., 2024a), FEVER (Beniwal et al., 2024), MzsRE (Wang et al., 2023), sometimes even different NMT model comparison as in mParaRel Fierro and Søgaard 2022), followed by using LLMs (as in MULTIGEN (Huang et al., 2024), DLAMA, BMIKE-53 (Nie et al., 2024) or MLaKE by Wei et al. 2024) and, in rare cases, by explicit rule-based inflection generation with the help of Unimorph-inflect (Anastasopoulos and Neubig, 2019) (such as in X-FACTR Jiang et al. 2020a). Some projects explicitly state that the whole dataset was manually checked after the translation phase (such as GeoMLAMA), others, including MLAMA, state that they covered a subset of languages and relations at the manual check round. Importantly, the procedure of translating datasets with NMT may differ in setups: while datasets like MzsRE translated the full sentences with Google Translate, the MLAMA creators translated only the relation templates (e.g. "[X] was born in [Y]" and then used the translations of the X and Y entities retrieved from WikiData directly, which ignores the grammaticality of the final sentences.

It is trivial to assume that translation LMs would increase the number of grammatical sentences in the dataset compared to template filling: we expect that based on the training data of both NMT systems and LLMs (which are usually fluent sentences). Consequently, we can see a gradual trend in switching from templates to sentence-level translation by NMT and then by LLMs, while the datasets like MLAMA are still being used in multilingual LLM evaluation. However, to our knowledge, there has been no direct comparison of templated (i.e. not necessarily grammatical or lexically accurate) to sentence-level translated versions of the datasets, hence the question of whether there is a real gain in using fluent sentences for multilingual LLM prompting, or templated sentences are already a sufficient baseline, still stands. With this paper, we attempt to fill this lacuna.

3 Method

We follow the general framing of factual knowledge that can be seen in (Elazar et al., 2021) or (Meng et al., 2022): a **fact** f_i is a triplet $\{s_i, r_i, o_i\}$, where subject s_i and **object** o_i are named entities (NE), and r_i is a type of relation between the two NEs. The structured fact then has to be formulated as a natural language sentence, which we call **verbalization** (following Khodja et al. 2025). We therefore say that an LLM "knows" the fact f_i if, prompted with a verbalization of a correct fact $\{s_i, r_i, o_i\}$ and of the incorrect triplets $\{s_i, r_i, o_*\}$ differing by object NE, the model assigns higher probability to the correct fact verbalization rather than to the incorrect triplets, which are called "**distractors**" (also following Khodja et al. 2025). In the following subsections, we will cover the fact, verbalization and distractor choice in more detail, as well as final evaluation.

3.1 Language, Fact and Distractor Choice

The initial MLAMA dataset comprises facts from 53 languages, which cover a broad typological, genealogical and geographical sample. At the same time, the authors note that the templates (not the final verbalizations) were checked manually only for three languages. In this work, since we focus on grammaticality and fluency, we narrow down the scope of languages and chose four Slavic languages, Russian (ru), Czech (cs), Ukrainian (uk) and Croatian (hr), and five non-Slavic languages, Spanish (es), simplified Chinese (zh), Vietnamese (vi), Indonesian (id) and Danish (da), in order to assess the fluency more attentively. A short information about the language families, scripts and resourcedness of the languages is provided in Appendix F.

The pool of facts in MLAMA comprised 41 relation types mostly from Wikidata (Vrandečić and Krötzsch, 2014). For the analysis, we choose a subset of 15 relation types from Wikidata only, which were unambiguous and had a distractor pool large enough. For an overview of the included and excluded relations, see Appendix B[2].

[2]It is necessary to note, that this limitation is not specific to our work: MLAMA and other highly multilingual factual datasets are not designed for direct cross-lingual comparisons. Very few datasets (like DLAMA by Keleg and Magdy, 2023) aim at balancing the datasets by the number and variety of

The pool of subject and object translations is retrieved from Wikidata directly. To enrich the distractor pool (and for the additional experiment in Section 5.2), we extract the aliases for all objects in question from Wikidata (i.e., for the entity denoting New York, we also retrieve its aliases like "NYC", "New York, NY", "The big Apple" etc.).

The distractors are chosen following the initial MLAMA Typed Querying setting: they need to be of the same entity type as the correct object. For the sake of compute load, we put the maximum boundaries on the number of distractors by sampling 50 entities from the total pool of objects for a particular relation.

3.2 Dataset Creation

We compare three types of verbalization of the same set of knowledge facts to see if sentence-level translation helps factual retrieval:

1. **Template** filling: the translated template formulations are borrowed from the MLAMA dataset as is and filled with NEs from Wikidata. This is the expected usage scenario for the MLAMA benchmark.

2. **GT**: The templates in English are first filled with English names for NEs and then translated with Google Translate.

3. **ChatGPT**: For the experiment with the Slavic languages, we provide the third verbalization: filled English templates (same as in GT) and the translations of the subject and object NEs (retrieved from Wikidata), and prompt GPT-4o-mini[3] to translate the full sentence using the provided NE translations. For each relation, we create few-shot instruction (20 examples) to show the model how we want it to use the Wikidata NE translations and how we expect it to use appropriate inflection of the relation verbalization. A few-shot prompt example is shown in Appendix A.

We intend to probe an LLM with a part of the verbalization that precedes the object NE. Thus, we split the generated verbalizations into the prompts and the expected endings (object NEs). The template verbalizations are split straightforwardly by cutting the "[Y]." suffix from the end of the sentence; the GT and ChatGPT verbalizations are split with the help of Stanza package (Qi et al., 2020) that searches for the (possibly inflected) object NEs in the sentences. Thus, for each fact, we get up to three different prompts and up to three forms of object NEs (sometimes the generated verbalizations or their parts may not differ); we consider all forms (both non-inflected and possibly inflected) of object NEs as correct ones. While the ChatGPT translation is expected to be consistent in word order of the prompts of the same kind, we cannot control that for Google Translate. Thus, we filter such facts from the dataset for which the GT verbalization differs significantly in word order (e.g. an object NE in the beginning of the sentence).

3.3 Experiment Setup

For each fact f_i, we concatenate the verbalization prompt v_i and either the set of correct object NE or 50 distractors sampled from the same language-relation subset of the dataset (in total, they form the candidate set). Next, we feed the model with such a set of prompts, calculate the log probabilities of tokens corresponding to object NEs, and rank the possible endings of the verbalization by these log probabilities. Then, for a given fact, its score is 1 if the object is ranked among the top n results, and 0 otherwise. We follow the principle of evaluation introduced in (Petroni et al., 2019), where it was called Precision at n; however, in our case the pool of adversarials is of a fixed size, thus, it is more appropriate to call it recall at n, $R_i@n$. To see the stability of the trends in verbalization difference, we generalize n up to 5 in some of our experiments, contrary to MLAMA score that only considers $R@1$.

We calculate the percentage of such hits for a given verbalization type (template, GT or ChatGPT) within a given language-relation subset and note it as $R@n$. We are interested in comparing the $R@n$ values within the same language between the verbalization types; thus, for the general results we will average the scores for the whole language dataset. We do not compare the scores between the languages since the sets of facts in different languages differ significantly. This limitation is not specific to our work: MLAMA and other highly multilingual factual datasets are not designed for direct cross-lingual comparisons. Very few datasets (such as DLAMA) aim to balance the datasets by the number and variety of facts in each language,

facts in each language, but such a dataset is significantly harder to create for more than 3-4 languages.

[3]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

but such a dataset is significantly harder to create for more than 3-4 languages.

We choose the LLaMA-2-7b model (Touvron et al., 2023) for our experiment, since this model is one of the few models that explicitly claims its functionality in all four Slavic languages in question. The experiments were run on single NVIDIA GeForce RTX 4090 GPU, the total time of the experiment was 23 hours.

4 Results

4.1 Slavic Languages

Our first experiment shows (see the left half of Table 1) that the verbalizations generated by Google Translate and by ChatGPT, in general, extract more facts from the LLM than the templated ones. Notably, the GT verbalizations also work considerably better than the ChatGPT-generated ones. In case of the higher-resourced languages (Czech and Russian), the increase in performance is up to 10% of the retrieved facts (under GT verbalization). The only exception is the language with the lowest resources, Croatian, where the template and GT scores oscillate around the same value with a decrease of ChatGPT performance. We also hypothesized if such an increase in performance is observed only in the $R@1$ setup, or the trend is steady across different top n values. In Appendix C we demonstrate that GT, followed by ChatGPT, has a steady delta across n up to 5; moreover, in the case of Croatian, the GT-generated verbalizations start extracting slightly more knowledge than the templated ones.

Another way of looking at the increase in factual retrieval depending on verbalization is to look at the distributions of the highest correct ranks for each fact. A demonstrative metric would be the mean value of a distribution which is given in the right half of Table 1: we see that, for the high-resourced languages, the GT verbalizations give an increase of the correct answers by one rank, and up to half of the rank for the low-resourced ones. Since the single mean value may not be an informative description of the whole distribution, the more detailed visualizations of the rank distributions are provided in Appendix C, showing similar outcomes.

The results above were focusing more on different verbalizations of the prompt beginnings. However, we should also remember that the MLAMA framework only counted the non-inflected forms of objects as the correct answers, while we mined the inflected object NE forms from the GT and ChatGPT verbalizations and allowed them as correct together with the non-inflected ones, and kept track of the ranks of both of them. We compare the rank differences between the inflected and non-inflected object forms by selecting 8 relation types where the object is expected to be inflected, and only such occurrences where exactly two forms of the ground truth object were found (assuming that the ground truth form of the templated sentence is the non-inflected form, and another one is the inflected one). Then, we subtract the rank of the non-inflected form from the inflected one and average such differences over language and verbalization type. The results in Table 2 show that, firstly, inflected forms are preferred over non-inflected ones in all setups (including the templates), and secondly, compared to templates, the delta in GT verbalization increases from 0.7 ranks (in Croatian) up to almost 3 ranks (in Russian). We therefore see that both the fluency of the beginning of the prompt and the correct inflection of the object NE contributes to increasing the retrievability of a fact from the model.

4.2 Non-Slavic Languages

In the experiment above, we restrict ourselves to four languages to manually create few-shot prompts for each relation and then check the resulting translations. Our results show that the biggest increase in retrieval is made with Google Translate verbalization, which does not require human editing. Thus, we decided to run comparison of two verbalizations, template and GT, on five more languages which belong to different language families and use various writing systems (still, we restricted ourselves to only such languages where we can judge the grammaticality of the final outputs post factum). Our sample included Spanish, simplified Chinese, Danish, Vietnamese and Indonesian. Notably, we also sampled a smaller number of relations (namely, eight) than for the first experiment: we selected only the relations that have object in the end of the sentence for all 5 languages (see motivation for that in Section 5.4).

Table 3 shows the results of this experiment[4]. The results show similar trends as for the Slavic languages: GT verbalization works as an upper bound for all languages, yielding up to 24% in $R@1$ score

[4]The experiments were run on single NVIDIA GeForce RTX 4090, the total time was 9 hours.

	R@1 ↑				Mean Rank ↓			
Verbalization	ru	cs	uk	hr	ru	cs	uk	hr
Template	0.512	0.579	0.488	0.707	4.91	4.17	5.12	2.94
GT	0.586	0.67	0.525	0.704	3.95	3.16	4.43	2.68
ChatGPT	0.545	0.615	0.492	0.653	4.65	3.53	5	3.02

Table 1: Main experiment results: comparison of the performance of three verbalization prompt types. The left columns represent $R@1$ metric (percentage of cases where the LLM ranked the correct object as the most probable; higher score is better retrieval). The right columns show mean value the most probable correct object rank for each fact (lower score is higher rank, therefore better ranking).

Verbalization	ru	cs	uk	hr
Template	5.80	8.49	7.86	8.31
GT	8.59	7.55	9.02	9.02
ChatGPT	7.54	7.28	9.07	8.54

Table 2: Average rank delta that shows how many ranks the inflected form of ground truth object is higher (closer to rank 1) than its non-inflected counterpart (higher number means bigger delta in favor of the inflected object).

and decreasing down to 5 ranks in average rank score. The only exception is Spanish, which remains stable in terms of both the $R@1$ and average rank score (for particular relation types, however, we see increase in $R@1$ up to 5%). The reasons for such an increase differ depending on language: for Danish and some relations in Spanish, it is proper choice of inflectional category, prepositions or articles; for languages with a narrow inflectional inventory like Vietnamese and Indonesian, the main effect is mostly explained due to more appropriate translations of the relations. With Chinese[5], most of gains in factual retrieval are explained by either preservation of the Latin spelling of the proper name, or by providing both the Chinese transliteration of the name and the initial Latin spelling in parentheses; we can consider it as one of the forms of explicitation described in Section 5.1. Overall, we consider this a good sign showing that the sentence-level translation of the dataset can curate a range of language-specific problems which are unrelated to each other.

[5]The overall quality of factual retrieval for Chinese is notably low. Our manual check of the results did not show any technical reason for such a poor performance. We hypothesise that this may be related to the poor tokenization quality of the texts written in Chinese characters (compared to Latin and Cyrillic scripts); for general discussion about the multilingual tokenization inequality and its effect on language performance refer to (Ahia et al., 2023) (which, however, does not include Chinese). We leave the analysis of that hypothesis for the future research.

5 Discussion

5.1 Qualitative Analysis

One of the most surprising observations in the metrics above was the significant prevalence of GT verbalizations over the ones generated by ChatGPT, despite even 20-shot prompting. The detailed analysis reveals the possible reason behind that: while ChatGPT was forced to follow the subject and object name translations from Wikidata labels and enforce the uniformity of relation formulations, Google Translate was not only free to choose the most common phrasing of the relation, but also to add clarifications for the subject entity type. Such additional information may help LLM to resolve the ambiguities (for instance, when an entity is a movie name which can also be used as a regular phrase, like "The Pianist") and prune it to the correct answer. This behavior of Google Translate can be called "explicitation" and is a well-known technique within translation practice (Murtisari, 2016). The type of explicitation we see in this case is called "pragmatic explicitation" (Klaudy, 1998) and is used to help the target audience of a different culture to better understand the text. Various types of pragmatic explicitation are seen in our dataset, starting with explicit definitions of the entity types, ending with punctuation such as brackets, which goes in line with (Ädel, 2006) that includes typographical markers into explicitation forms. Notably, even such minor additions as brackets can increase the rank of the retrieved entity to the top n choices.

Another notable observation was comparison of two opposite relations: "[X] is the capital of [Y]" and "The capital of [X] is [Y]" and its verbalizations, in Czech. In both cases, the template translation of the relation was completely wrong due to ambiguity of the term "capital" in English: it was translated as "kapitál" (as "capital goods"), not as correct "hlavní město" (as "capital city"). Interestingly, the $R@1$ value for the relation "The capital

	R@1 ↑					Mean Rank ↓				
Verbalization	es	zh	vi	id	da	es	zh	vi	id	da
Template	0.725	0.02	0.587	0.547	0.568	2.64	19.98	4.1	3.7	4.16
GT	0.725	0.059	0.765	0.786	0.64	2.61	15.68	2.59	2.12	3.84

Table 3: $R@1$ and average rank scores for five additional languages. The notation is the same as in Table 1.

of [X] is [Y]" (i.e. when the country entity was provided and the model had to guess the capital entity) increased from 0.52 (template verbalization) to 0.89 (GT verbalization), while the opposite relation (where the model had to guess the country given the capital) increased from 0.79 (template) to 0.83 (GT). The disparity in factual retrieval of the opposite relations is known as the "reversal curse" in the LLM evaluation (Berglund et al., 2023) and is observed frequently, and if we were to compare only the template verbalization of the two relations, we would get such an asymmetry; however, we see that the correct translation of the relation "curates" such a disparity and decreases the gap almost to zero. This example also highlights the necessity of sentence-level translation of prompts, since this minimizes the risk of the wrong choice of relation disambiguation in the target language.

5.2 Named Entity Aliases: Subjects and Objects

As noted in Section 3.1, both subject and object NEs have different variants of phrasing in Wikidata. For each entity, there is a default way of phrasing it and the alternatives called "aliases". The MLAMA benchmark did not take into account this plurality of NE naming; in other benchmarks, we only know of X-FACTR (Jiang et al., 2020a) counting any of the object aliases as the correct answers, while the subject aliases, to our knowledge, remain not analyzed. The default Wikidata phrasing of the entities is language- and culture-dependent and sometimes can be arguable. For example, the default phrasings of the personal names in the Russian Wikidata are provided in the form of "Surname, First Name (Patronymic if applicable)": "Tolstoi, Lev Nikolaevich", which is formally grammatical but has a strong "bureaucratese" connotation, whereas a formal but more neutral way of addressing will be "First name (Patronymic) Surname": "Lev Nikolaevich Tolstoi". As a more general case, we can think of the translation VS preservation of the movie or book titles: in the Czech and Croatian versions of Wikidata, the translated version is slightly preferred as the default phrasing.

Our experiment is designed in such a way that in ChatGPT verbalization, the translation of the subject and object NEs was controlled (by providing the default translations from Wikidata into target languages), while in GT setting the system chose the most probable translation of the English NE itself. Thus, we see multiple examples of the choice of alternative namings of both subject and object NEs by Google Translate over all relations and all four languages. For example, for the Russian personal name translation, GT chooses more neutral way of addressing.

The analysis of variation of different subject phrasing in GT versus two other verbalization does not show significant effect. The neutral phrasing of the Russian personal names does not yield much for retrieval (compared to formal phrasing in templates and ChatGPT, the increase is from 0.01 to 0.06 score depending on relation type). The preservation of the English book or movie name by GT (instead of translation suggested by Wikidata) also does not show consistent change, with anecdotal evidence of increasing the retrieval quality for Czech and Croatian and decreasing it for Ukrainian. In most cases such a variation goes in hand with the explicitation artifacts described above, which seems to have a bigger effect on fact retrieval.

Evaluating the variation of all possible subject phrasings would have been too computationally-consuming (as we would have to create prompts for all possible aliases of the prompt for the correct inflection of the relation). With the object NE aliases, we make an experiment by including all possible aliases into the pool of the correct answers (we do not inflect the objects for the same reason of preserving compute by not generating the new sentences; on the other hand, we include the English NE labels to include the probability of activating the representations of the most resourced language). We re-run the evaluation for all languages and relations given this wider pool of correct answers; the results on $R@1$ score can be found in Appendix E. We see that while all scores increase, the trend

remains the same, with GT being the best verbalization, followed by ChatGPT and templates.

Our comparison shows that, despite variation in object and subject NE, the more important factors in factual retrieval are the overall fluency of the prompt and, in some cases, various explicitation techniques.

5.3 Language Resourcedness

One of the motivating ideas of our experiments was the expectation that the increase in factual retrieval will be more noticeable for the lower-resourced languages, while the high-resourced languages could manage retrieval well even under distorted input. In our sample, the most resourced language is Russian among Slavic languages, and Spanish among non-Slavic languages (0.13% of LLaMA-2 training data each). Ukrainian and Vietnamese are seen less (0.07% and 0.08% training data, respectively), while the rest of the languages (Czech, Croatian, Indonesian, Danish) are seen in from 0.03% to 0.01% training data.

In reality, we did not observe any clear trend: the biggest increase in factual retrieval was for Russian and Czech; the improvement for Ukrainian was smaller, while Croatian showed comparable performance for templated and GT verbalizations. If we look at the sample of languages from our second experiment, the trends become even less clear: Chinese and Spanish are equal to Russian in Llama-2 training data (0.13% each), and show completely different patterns in factual retrieval. Similarly, Danish and Croatian are almost equally low-resourced in the training data (0.02% and 0.01%, respectively), yet Danish benefits from sentence-level translation considerably more.

5.4 Open Questions

1. How to measure grammaticality? In our experiments, we assume beforehand and then check manually the grammaticality of three types of verbalizations. However, this approach is too time-consuming and is hardly scalable. To approximate the grammaticality of sentences, we computed the log probabilities of different verbalizations for the same fact; this metric was not consistent in terms of showing the fluent sentences. Neither did help surface-based metrics such as chrf (Popović, 2015) that we applied to compare template sentences against GT or ChatGPT verbalizations. Additionally, we applied the grammar error correction (GEC) systems for Russian and Czech such as (Rothe et al., 2021) to see whether it would detect grammatical inconsistencies in templated sentences. We observed that the GEC systems mostly targeted spelling mistakes within the words rather than grammatical agreement, therefore such systems also cannot help us with the grammaticality evaluation. Our observation goes in line with the overviews of the GEC systems for languages other than English (such as in (Volodina et al., 2023).

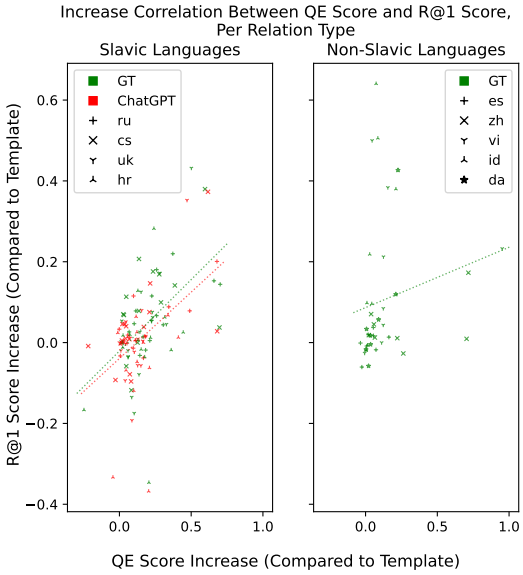

Figure 2: Correlation between increase in QE scores (x-axis) and increase in $R@1$ performance (y-axis) of the GT and ChatGPT verbalizations compared to templates. The left subplot shows statistics for the first experiment (Slavic languages), the right - for the second experiment (non-Slavic languages). Each point represents the averaged value over a particular language, relation and verbalization type. The color represents the verbalization type; the shape of the point stands for the language. The dotted lines show linear regressions fit by all data points at each subgraph for a corresponding type of verbalization.

We can also hypothesise that a quality estimation score can be an indirect estimation (QE) of the grammaticality of the sentence (although, according to classical framing of the MT task, it estimates accuracy and fluency of the translation at the same time). Indeed, the scores that we obtained with the COMET model[6] showed a considerable increase of QE scores (up to 9% for Slavic and up to 23% for non-Slavic languages), if we compare the templated translations to GT or ChatGPT outputs. Moreover, the increase of QE scores shows positive correlation with the increase of our target metrics,

[6] https://huggingface.co/Unbabel/wmt20-comet-qe-da

for instance, $R@1$ (up to 0.8 Pearson's score). This is illustrated in Figure 2, and the detailed information about the observed correlations can be found in Appendix D.

Thus, to make grammaticality evaluation scalable, we need to operationalize the notion of grammaticality and create the respective metric for it.

2. How to prompt models multilingually? We follow the spirit of the MLAMA probe in our experiment design: MLAMA benchmark was created for encoder models, whose main training objectives and inference functions was masked token prediction. Thus, the templated sentences were fed to mBERT with the object token substituted by "[MASK]" sequences. Analogously, since the pre-training objective and inference function for decoder-only models is next token prediction, we formulate the prompts in such a way that the prompt verbalization would contain the full name of the subject NE and the formulation of the relation, while the comparison will be done for the log probabilities of the tokens making up object NEs which are placed in the end of the sentences (e.g., a verbalization would look like "Leo Tolstoi was born in" + object entities). However, the languages differ by the so-called basic word order (the most unmarked placing of subject, verb and object in declarative sentences), which is formalized as SVO, VSO etc. (for more information, see the WALS project by Dryer et al. 2024). Noticeably, our prompting strategy will work for languages which are O-final; however, there is a large number of languages that are V-final (for example, Turkish, Hindi and Japanese), where the explication of the relation (mostly formulated through a verb phrase) is put after the object. If we apply the existing MLAMA benchmark to such languages for a decoder model in a straightforward way (by splitting the sentence at the place of object), this will not provide the model with information about the relation. Alternatively, we can come up with another prompt schema: showing the whole sentence with a masked object and then ask the model to name the object (e.g. "Leo Tolstoi was born in [Y]. Y:" This may work well in the instruction-tuned versions of the models (especially after few-shot examples), but with this setup we will not be able to disentangle the parametric factual knowledge from the alighment skills learnt through stages like SFT and RLHF. Therefore, we incline towards prompting the model in the way we did (finish the sentence with missing object), but extension of that framework to verb-final languages from the MLAMA pool would require additional curation of the verbalizations so that the objects would be put in the end of the sentence. In most cases, this is possible even for languages with SOV and OSV word orders, as such languages do allow some structures with an object in the end (usually some types of highlighting of one of the constituents), but for this task, careful few-shot prompting through ChatGPT and post-checking by fluent speakers will be required.

6 Conclusions

This work analyzes different types of prompting for multilingual factual knowledge evaluation of an LLM depending on whether the prompts are formulated in a fluent manner (translated by an NMT or by an LLM as a whole sentence) or ignoring the grammatical and lexical consistency (filling of the fixed templates). We compare different verbalization for four Slavic and five (typologically and graphically diverse) non-Slavic languages, and show, to our knowledge, the first evidence for an intuitive assumption that sentence-level translations help retrieving more facts from the model than templated translations. This is an indication that cross-lingual factual retrieval may be underestimated in experiments based on MLAMA or derivative datasets that are template-based. With these results, we encourage the community to use properly translated datasets for higher and more interpretable results of multilingual factual knowledge prompting.

Limitations

Our work has several limitations:

1. Small samples over different dimensions. The analysis was done for four languages from the same group within the Indo-European family, plus five typologically diverse languages. This narrow focus served to manually create/verify parts of the datasets and to do qualitative analysis. The relation sample represents less than a half of initial relations from MLAMA. This happened for several reasons: some relations were too ambiguous, too broad in terms of entity types included, had too few object entities, or their template verbalization did not have an object in the end of the prompt. We also tested that on a single LLM since this was the only popular model that explicitly states that it was trained on all languages from our sample.

2. Lack of control in GT. We could not control the exact content of the sentences that were generated by the NMT system (Google Translate), which resulted in artifacts such as explicitation. This also seems to be the reason for the highest performance of GT compared to a more controlled ChatGPT-generated translations. Yet, even the latter shows increase in the factual retrieval, which is shown both by $R@1$ metric and by ranking distribution of the correct answers. On top of that, due to variation in word order and translation choice of the entities in GT, we had to filter out the facts that were inconsistent with a format of a prompt, which decreased the number of data points in the resulting dataset.

3. Distractor choice. We sampled the distractors in a pseudo-random reproducible manner (with the help of hashing). However, there is an open discussion on what should be optimal distractor choice (Khodja et al., 2025). Still, we think that our distractor samples (50) are big enough to ensure variety within them, especially since the metric $R@n$ is considering only up to 10% of the first ranks of the candidate set.

4. Fact choice. The problem inherited from MLAMA is the imbalance of facts between languages in the initial dataset (up to 10 times between the highest- and the lowest-resourced languages). Therefore, in our final factual sample, Croatian was the smallest with less than 2000 facts, compared to Russian, Czech and Ukrainian with approximately 6600, 3900 and 4100 facts, respectively. Additionally, the set of entities and relations is West- and specifically Anglo-centric, which was addressed in the datasets like (Keleg and Magdy, 2023). Still, we believe that our comparison that was done on several thousand facts for each language allows us to highlight the importance of the proper translation of the sentences even for the imbalanced benchmarks.

Acknowledgments

This research was funded by NCCR Evolving Language, Swiss National Science Foundation Agreement 51NF40_180888. We also thank the following people for helping with the preparation of the few-shot prompts and qualitative analysis of the larger sample of languages: Yulia Alpaieva (Charles University) for Ukrainian, Michelle Wastl (UZH) for Croatian, Nam Luu (University of the Basque Country) for Vietnamese, Sophia Conrad (UZH) for Danish, Polina Nalsedskova (HSE University) for Indonesian.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the Limits of Low-Resource Morphological Inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual Editing in Multilingual Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint*. Version Number: 4.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the Center of the Knowledge Neurons: Discoveries of Language-Independent Knowledge Neurons and Degenerate Knowledge Neurons. *arXiv preprint*. Version Number: 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. Version Number: 2.

Matthew Dryer, Martin Haspelmath, Matthew Dryer, and Martin Haspelmath. 2024. The World Atlas of Language Structures Online.

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, Arslen 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)*, pages 3451–3452, Miyazaki, Japan. European Language Resources Association (ELRA).

Constanza Fierro and Anders Søgaard. 2022. Factual Consistency of Multilingual Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly. 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 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024. 1+1>2: Can Large Language Models Serve as Cross-Lingual Knowledge Aggregators? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13394–13412, Miami, Florida, USA. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. 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.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Hichem Ammar Khodja, Abderrahmane Ait gueni ssaid, Frederic Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2025. Factual Knowledge Assessment of Language Models Using Distractors. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8043–8056, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kinga Klaudy. 1998. Explication. In M. Baker and G. Saldhana, editors, *Encyclopedia of Translation Studies*, pages 80–85.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *arXiv preprint*. Version Number: 5.

Elisabet Murtisari. 2016. Explicitation in Translation Studies: The journey of an elusive concept. *The International Journal of Translation and Interpreting Research*.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. 2024. BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning. *arXiv preprint*. Version Number: 2.

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, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan Iv, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can Language Models be Biomedical Knowledge Bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. Version Number: 2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-Lingual Knowledge Editing in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023. Knowledge Editing for Large Language Models: A Survey. *arXiv preprint*. Version Number: 4.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024b. Retrieval-Augmented Multilingual Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. MLaKE: Multilingual Knowledge Editing Benchmark for Large Language Models. *arXiv preprint*. Version Number: 3.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language Anisotropic Cross-Lingual Model Editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking Knowledge Boundary for Large Language Models: A Different Perspective on Model Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Annelie Ädel. 2006. *Metadiscourse in L1 and L2 English*, volume 24 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, Amsterdam.

A Few-Shot Prompt Example

Below you can find an example of the few-shot prompt used to generate the ChatGPT verbalization of a fact. A different prompt was created for each language and each relation type. For the sake of brevity, we show here 5 shots of a prompt for Czech language and relation P19 (birthplace of a person); the rest of the 20 prompts in total were formatted the same way. At inference, the model is provided with "Source sentence", "Subject translation" and "Object translation" fields, and is expected to generate the translated sentence only after the word "Translation: " (see the last paragraph on the page).

Note that the fields "Subject translation" and "Object translation" are extracted from Wikidata and then are forced to be used as translations of the corresponding entities; yet they can be inflected if necessary: n this particular relation, the object entities are put in locative case, e.g. "Lucembursko" (Nominative) -> "Lucembursku" (Locative). Apart from that, the verb is also inflected differently ("narodila se" or "narodil se" depending on gender of the subject).

```
You are a professional English-Czech translator. You are given English sentences
about subjects and objects. You are also given translations of subjects and objects
separately. You need to translate full sentences to Czech. When translating, you
have to use the translated subjects and objects. Pay special attention
to grammatical agreement between the words in the translated sentences.
When translating, follow the examples:

    Source sentence: Cunigunde of Luxembourg was born in Luxembourg .
    Subject translation: Kunhuta Lucemburská
    Object translation: Lucembursko
    Translation: Kunhuta Lucemburská se narodila v Lucembursku.

    Source sentence: Karel Schwarzenberg was born in Prague .
    Subject translation: Karel Schwarzenberg
    Object translation: Praha
    Translation: Karel Schwarzenberg se narodil v Praze.

    Source sentence: William Carlos Williams was born in Rutherford .
    Subject translation: William Carlos Williams
    Object translation: Rutherford (New Jersey, U. S.)
Translation: William Carlos Williams se narodil v Rutherfordu (New Jersey, U. S.).

    Source sentence: Marion Davies was born in Brooklyn .
    Subject translation: Marion Daviesová
    Object translation: Brooklyn
    Translation: Marion Daviesová se narodila v Brooklynu.

    Source sentence: Peter Mark Roget was born in London .
    Subject translation: Peter Roget
    Object translation: Londýn
    Translation: Peter Roget se narodil v Londýně.

    ...

    Source sentence: Theodore the Studite was born in Constantinople .
    Subject translation: Theodoros Studijský
    Object translation: Konstantinopol
    Translation:
```

B Relations Overview

Relation ID	en Template	ru	cs	uk	hr	es	zh	vi	id	da
P101	[X] works in the field of [Y] .	605	396	396	226	–	–	–	–	–
P103	The native language of [X] is [Y] .	736	458	445	231	882	554	269_{36}	334	737
P108	[X] works for [Y] .	91	51	52	18_{28}	149	162_{46}	46_{38}	52_{34}	92
P127	[X] is owned by [Y] .	402	246	258	84	–	–	–	–	–
P1376	[X] is the capital of [Y] .	217	214	216	174	–	–	–	–	–
P159	The headquarter of [X] is in [Y] .	450	240	303	84	483	465	93	221	259
P19	[X] was born in [Y] .	493	248	193	103	919	272	86	194	597
P20	[X] died in [Y] .	679	418	342	158	–	–	–	–	–
P36	The capital of [X] is [Y] .	614	453	617	294	647	538	361	428	439
P364	The original language of [X] is [Y] .	373	181_{43}	212	64_{33}	394	319	71_{24}	183	228_{44}
P407	[X] was written in [Y] .	583	355	404	163_{42}	702	223	193	287	342
P449	[X] was originally aired on [Y] .	242_{28}	99_{27}	96_{18}	39_{18}	–	–	–	–	–
P463	[X] is a member of [Y] .	151_{45}	54_{40}	130_{30}	102_{29}	–	–	–	–	–
P495	[X] was created in [Y] .	414	205	240	81	–	–	–	–	–
P740	[X] was founded in [Y] .	416	256	214	88	782	275	78	160	233

Table 4: Overview of relations used in the main experiment.

Table 4 shows statistics about the relation types and the sizes of datasets that make up these relations used in the experiments. The "Relation ID" column represents the Wikidata ID, "en Template" column shows the English template that was extracted from the English version of MLAMA (it was then used as a ground for GT and ChatGPT sentence-level translations into the target languages). The columns with language IDs show the number of facts in each dataset that were used in the experiment (i.e. after translation and filtering out of the GT verbalizations which had a wrong word order). By default, the number of distractors that were used for each relation type is equal to 50. In case of difference, this number (which is equal to the whole pool of other entities for a given relation) is provided in subscript for the corresponding relation type and language.

For the first experiment with the Slavic languages, fifteen relations out of the whole pool were used. These fifteen relations comprise only one third of the relations from the initial MLAMA dataset. The reasons for filtering out other relations were the following:

1. The prompt formulation in at least one of the target languages did not have object in the end of the sentence, for example, relation P106 is formulated as "[X] is a[Y] by profession." and similarly in other templates. The reason for excluding such phrasings is provided in Section 5.4.

2. The relation has too few unique object NEs. For example, relation P413 "[X] plays in [Y] position." has fewer than 10 unique objects, which would make the distractor pool too narrow and automatically increase the metrics.

3. Relations that are too broad and/or ambiguous. For instance, relation P527 "[X] consists of [Y]" includes chemical elements ("water consists of hydrogen"), deities ("Dashavatara consists of Rama"), political entities ("G20 consists of Italy") and so on. Moreover, if another correct object will be in the distractor pool (for instance, "Germany" for the prompt "G20 consists of "), it would not be scored.

4. Relations that are poorly formulated in English. Examples include P1001 "[X] is a legal term in [Y]." (which rather means the objects that apply to a particular jurisdiction, such as a parlament or a national flag of a particular county), or P140 "[X] is affiliated with the [Y] religion." (which either means the beliefs of a particular person, or the religion to which a particular temple (church/mosque/gurudwara/etc.) is dedicated).

For the second experiment, due to variation in the word order in typologically differences, we used only eight relations from the pool of the first experiment, such that the objects in the templates are sentence-final (see reason 1 above).

C Detailed Statistics of the Experiment With Slavic Languages

In Figure 3, we show the trends in the $R@n$ metric scores for each language depending on the value of n. We see that the scores for each verbalization increase consistently, showing the same trend of GT > ChatGPT > template verbalization for each language except Croatian. Thus, we see that the superiority of the sentence-level translation verbalizations is robust against the choice of n.

In Figure 4, we show the box plots that represent the distributions of the highest ranks of the correct object for each verbalization in each language. This gives us a broader view at the distributions compared to a single average rank metric. We see that, apart from being closer to one on average, the ranks of the sentence-level translated verbalizations (especially GT) have a narrower spread compared to templated ones, which makes their performance more stable.

For both graphs (as well as any other score in the paper), we show scores from a single run since we ensured the reproducibility of the ranking by making reproducible sets of correct objects and distractors (by sampling the entities by their hash IDs).

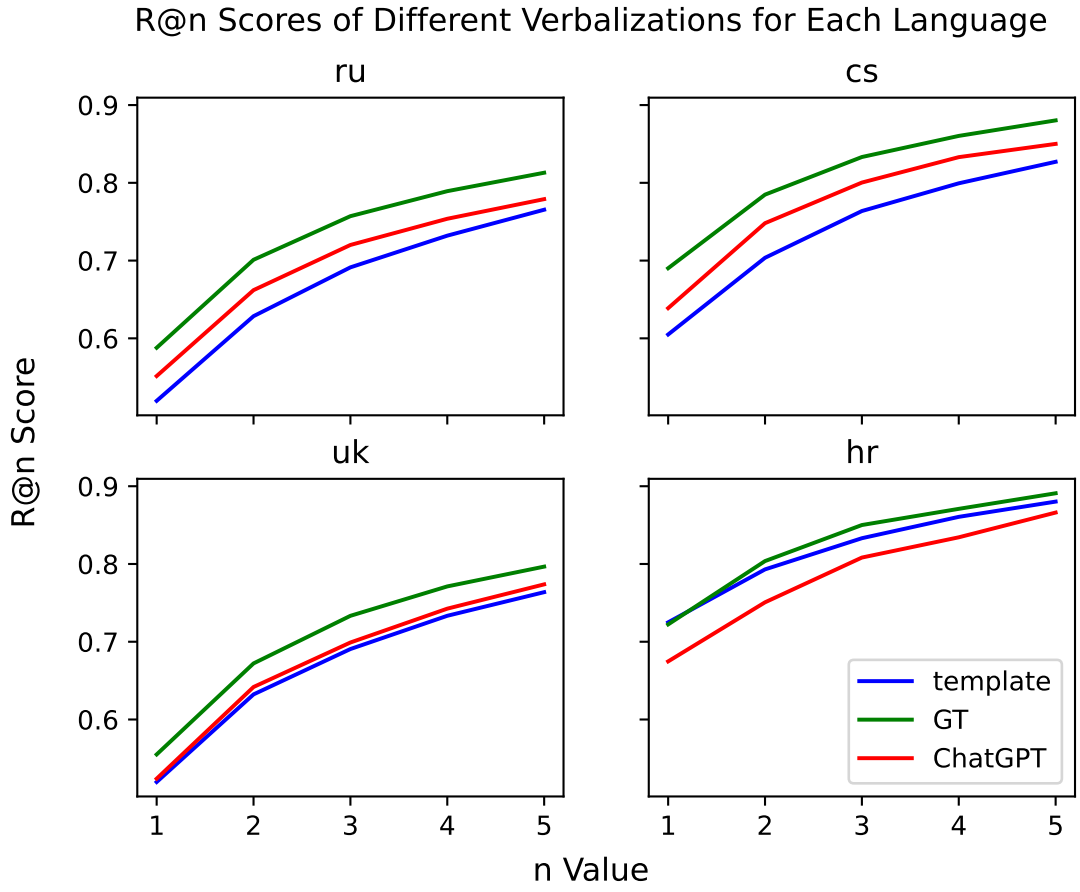

Figure 3: Comparison of $R@n$ scores depending on n. Higher score shows better retrieval.

D Grammaticality Estimation

The only consistent predictor for grammaticality estimation that we could find were the QE metrics. We demonstrate it with the increases in GT and ChatGPT performance (compared to templated translation) both in QE score and in our target $R@1$ metric. In the Table 5, we show the scores of the increases in QE and their correlations to $R@1$, averaged by language. The correlation is also demonstratively shown on a more granular level of the language-relation pairs in Figure 2.

We also made additional evaluation of grammaticality of the prompts in a more controlled manner by concentrating on specific relation types. Namely, we analyzed the quality of handling the grammatical

Metric	Verbalization	ru	cs	uk	hr	es	zh	vi	id	da
Δ	GT	0.073	0.091	0.038	-0.003	-0.0	0.039	0.179	0.239	0.072
	ChatGPT	0.032	0.036	0.005	-0.054	N/A	N/A	N/A	N/A	N/A
r	GT	0.608*	0.498	0.788*	0.26	0.436	0.408	0.135	0.592	0.838*
	ChatGPT	0.78*	0.613*	0.741*	0.206	N/A	N/A	N/A	N/A	N/A

Table 5: Detailed statistics on QE score difference and its correlation with $R@1$ metric increase, averaged for every language. The QE score difference is calculated by subtracting the QE score of the template verbalization from either GT or ChatGPT verbalization. These scores are shown in the two upper rows of the table (under the Δ label). In an analogous manner (by subtracting the templated score from the scores from GT or ChatGPT verbalization) we compute the increase in the $R@1$ metric, and calculate the resulting Pearson's correlation between the increase of QE and increase of $R@1$ (to compute the correlation within each language, we firstly average the differences for each relation type). The correlation coefficients are represented in the two lower rows in the table (under the r label); the significant correlation ($p < 0.05$) is marked with asterisk.

Verbalization	ru		cs		uk		hr	
	%(F)	R@1(F)	%(F)	R@1(F)	%(F)	R@1(F)	%(F)	R@1(F)
Template	0.0	0.373	0.0	0.375	0.0	0.311	0.0	0.545
GT	80.7	0.38	89.8	0.591	90.2	0.246	84.8	0.515
ChatGPT	93.3	0.447	97.7	0.466	96.7	0.361	97.0	0.485

Table 6: Statistics on subset of relations P19 and P20 ("[X] was born in [Y]" and "[X] died in [Y]"), where X is a woman (and therefore the verb has to be in feminine gender). The statistics are provided for each language separately. The "%F" column demonstrates the percentage of the verbalizations with female subject where the feminine gender was used correctly; the "R@1(F)" column shows the R@1 score for the subset of the facts with the female subjects.

gender in Slavic languages. For that, we selected two relations, "[X] was born in [Y]" and "[X] died in [Y]", (P19 and P20, respectively). The verbs in the past tense are conjugated by gender (masculine, feminine or neutral). The templated verbalizations always used the masculine form of verb, irrespective of the gender of the person [X]. Therefore we concentrated on the increase in grammatical coherence of the verbalizations verbalizations for the persons of the female gender. For that, we extracted information about their sex/gender from Wikidata, and selected the subset of facts correlating only to women (selecting values "female" or "transgender female"). Then, we counted the percentage of feminine verb forms for each verbalization in each language, as well as the R@1 score for these (only female-related) subsets of facts. The results of this analysis are shown in Table 6. We can see that, firstly, the GT and ChatGPT verbalizations became predominantly accurate in terms of the verb form usage; secondly, for most of the cases, these verbalizations increased the quality of the factual extraction of the corresponding facts. Interestingly, for the subset of facts that are related to women, for most languages, the bigger increase in grammaticality (and in factual retrieval) is connected to ChatGPT verbalizations, not GT ones.

E Additional Experiments

Table 7 shows the main statistics for the experiment where we allowed Wikidata aliases of objects as correct answers[7]. They show increase in both $R@1$ score and average rank; however, the ranking of three types of verbalizations and the delta between them remains almost unchanged. This is an evidence of an importance of the whole sentence translation that cannot be approximated by mere expansion of the correct object phrasing pool.

F Language Overview

The languages chosen for the experiments (sorted by the mass in training data) are listed below:

[7]The experiments were run on single NVIDIA GeForce RTX 4090, the total time was 25 hours.

	R@1 ↑				Mean Rank ↓			
Verbalization	ru	cs	uk	hr	ru	cs	uk	hr
Template	0.57	0.63	0.539	0.736	3.79	3.18	4.04	2.43
GT	0.633	0.709	0.583	0.736	3.27	2.49	3.52	2.29
ChatGPT	0.59	0.67	0.54	0.693	3.76	2.84	4.01	2.54

Table 7: $R@1$ and average rank scores when aliases are allowed as correct answers. The trends remain the same as in the main results.

First experiment, Slavic languages (subgroups of the Slavic group are provided below):

1. **Russian**, code: "ru", East Slavic subgroup, Cyrillic script, 0.13% of pre-training data;
2. **Ukrainian**, code: "uk", East Slavic subgroup, Cyrillic script, 0.07% of pre-training data;
3. **Czech**, code: "cs", West Slavic subgroup, Latin script, 0.03% of pre-training data;
4. **Croatian**, code: "hr", South Slavic subgroup, Latin script, 0.01% of pre-training data.

Second experiment, non-Slavic languages:

1. **Spanish**, code: "es", Indo-European (Romance) family, Latin script, 0.13% of pre-training data;
2. **Chinese**, code: "zh", Sino-Tibetan family, Chinese (simplified) script, 0.13% of pre-training data;
3. **Vietnamese**, code: "vi", Austroasiatic family, Latin script, 0.08% of pre-training data;
4. **Indonesian**, code: "id", Austronesian family, Latin script, 0.03% of pre-training data;
5. **Danish**, code: "da", Indo-European (Germanic) family, Latin script, 0.02% of pre-training data;

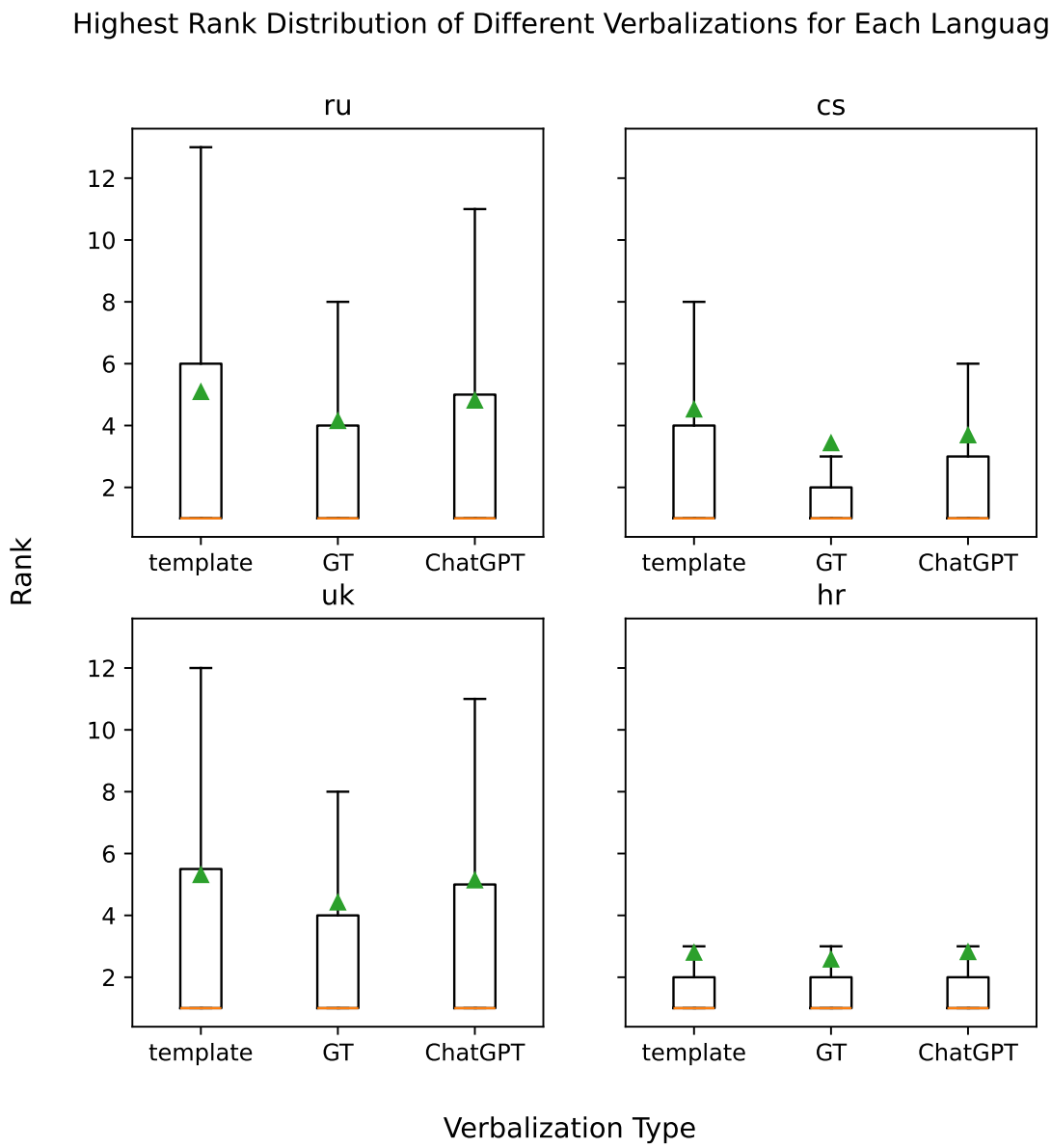

Figure 4: Distribution of the highest correct ranks for each verbalization in each language; lower score is better. For clarity, the box plots do not show outliers, which explain the skew of the average ranks (shown as green triangles) higher than the third quartiles.