Factual Knowledge Assessment of Language Models Using Distractors

Hichem Ammar Khodja[1,2], Abderrahmane Ait Gueni Ssaid [1], Frédéric Béchet[2,3], Quentin Brabant[1], Alexis Nasr[2], Gwénolé Lecorvé[1]
[1]Orange - *Lannion, France*,
[2]Aix Marseille Université, CNRS, LIS, UMR 7020 - *Marseille, France*,
[3]International Laboratory on Learning Systems (ILLS - IRL2020 CNRS)
Correspondence: {hichem.ammarkhodja, quentin.brabant, gwenole.lecorve}@orange.com,
aaitguenissaid@gmail.com, {frederic.bechet, alexis.nasr}@lis-lab.fr

Abstract

Language models encode extensive factual knowledge within their parameters. The accurate assessment of this knowledge is crucial for understanding and improving these models. In the literature, factual knowledge assessment often relies on cloze sentences, which can lead to erroneous conclusions due to the complexity of natural language (out-of-subject continuations, the existence of many correct answers and the several ways of expressing them). In this paper, we introduce a new interpretable knowledge assessment method that mitigates these issues by leveraging distractors—incorrect but plausible alternatives to the correct answer. We propose several strategies for retrieving distractors and determine the most effective one through experimentation. Our method is evaluated against existing approaches, demonstrating solid alignment with human judgment and stronger robustness to verbalization artifacts. The code and data to reproduce our experiments are available on GitHub*.

1 Introduction

Language Models (LMs) encode vast amounts of factual knowledge within their parameters (Petroni et al., 2019; Roberts et al., 2020). Assessing this knowledge is crucial for understanding the capabilities and limitations of these models (Kassner and Schütze, 2020), as well as for improving their performance in various applications (Jiang et al., 2020). Despite extensive research in this area, defining a proper measure of factual knowledge within LMs remains an open problem.

Looking at the current literature, facts are generally represented as relations between entities in the form of $(\textit{subject}, \textit{relation}, \textit{object})$ triples, or (s, r, o) for short. Examples include $(\textit{France}, \textit{capital}, \textit{Paris})$ or $(\textit{Germany}, \textit{shares a border with}, \textit{Switzerland})$. Assessing the knowledge of some fact (s, r, o) by an LM commonly relies on verbalizations of the pair (s, r) as a cloze sentence, which can have the form of a question (e.g., $(\textit{France}, \textit{capital}) \rightarrow$ "*What is the capital of France?* ____") or an incomplete declarative sentence (e.g., "*The capital of France is* ____"). These sentences are concise, target specifically the tested fact, and their expected continuation is a reference to the object o.

Due to the complexity of natural language, cloze sentences can be continued with an Out-Of-Subject (OOS) continuation that is linguistically correct but not informative regarding the fact under study (e.g., "*The capital of France is a city of contrasts.*"). Moreover, non-functional relations can associate many correct objects to (s, r). Thus, simply verifying that the LM generates "*Poland*" given "*Germany shares a border with*" is insufficient to determine whether it knows the fact $(\textit{Germany}, \textit{shares a border with}, \textit{Switzerland})$.

In this paper, we introduce a new knowledge assessment method that mitigates these issues by leveraging *distractors*, i.e., , incorrect but plausible alternatives to the correct answer. For a given LM, the probability of generating the correct answer from a cloze sentence is expected to be higher than the probability of generating a distractor. This approach is promising because restricting assessment to comparisons between entities inherently eliminates the issue of OOS continuations. Relying on Wikidata to collect facts and entities, we show that our distractor-based knowledge measure correlates with human knowledge assessment, among several measures from the literature, and is the least sensitive to verbalization errors, which are common when using cloze sentences.

The rest of this document is organized as follows: the state of the art of knowledge assessment methods is reviewed in Section 2 ; then, our distractor-based knowledge measure is introduced in Section 3 ; different distractor retrieval strategies are compared, and our knowledge measure

[*] github.com/Orange-OpenSource/DistFactAssessLM

Proceedings of the 31st International Conference on Computational Linguistics, pages 8043–8056
January 19–24, 2025. ©2025 Association for Computational Linguistics

is compared to other measures in the domain in Section 4.

2 Related Work

Assessing the knowledge of LMs is an active research area. While part of it includes the study of *linguistic* knowledge embedded in transformers by probing their latent representations (Hewitt and Manning, 2019; Tenney et al., 2019; Jawahar et al., 2019), our focus in on assessing *factual* knowledge.

As formulated in Petroni et al. (2019), most approaches represent facts as RDF triples from knowledge graphs. Knowledge is then assessed using predefined templates to convert triples to natural language (e.g., "*The capital of [SUBJECT] is [OBJECT]*"). Although Jiang et al. (2020) highlighted the challenges of this approach—notably regarding the large number of ways a fact can be verbalized, many methods have been proposed.

One category of methods involves generating multiple continuations for a cloze sentence using a decoding strategy such as beam search (Wiseman and Rush, 2016) or contrastive search (Li et al., 2023). The generated continuations are then analyzed using well-known metrics/techniques from the domain: **BERT-score** (Zhang et al., 2020), **ROUGE-L** (Lin, 2004), and **LLM-as-a-judge** methods (Sun et al., 2024; Zheng et al., 2023). These methods have the disadvantage of being vulnerable to OOS continuations because of the possibility that all or a large portion of the generated continuations are OOS, giving little information on the LM's knowledge for the tested fact.

A more direct approach would be to use, as a knowledge measure, the **conditional probability** of the right answer as the continuation of a given cloze sentence. However, interpreting this probability in terms of knowledge requires to set some threshold above which the fact is considered as known. Also, probabilities cannot be compared across different models because of the biases introduced by the training dataset. For example, an LM trained on a QA dataset is more likely to generate the correct answer after a question compared to an LM trained on CommonCrawl, not because it "knows" the fact better, but because it was conditioned to answer questions with entities. A more sophisticated knowledge measure based on conditional probability is **KaRR** Dong et al. (2023), which estimates the ratio between the probability to generate the correct answer given the LM's distribution and by pure chance. However, it suffers from the same limitations as using the conditional probability of the correct answer. This metric ranges over $[0; \infty[$, which further impedes interpretability.

Finally, **Precision@**n, Petroni et al. (2019) checks whether the correct answer is in the top-k most probable continuations to the cloze sentence. This has the disadvantage of being limited to single-token answers, because of the explosion of the number of continuations for answers encoded in more than one token.

As a consequence, our work aims to propose a measure that addresses the various gaps previously mentioned and summarized in Table 1. Indeed, our distractor-based approach is easily interpretable, free from the problems of OOS continuations. Furthermore, it allows comparison across different models, answers with multiple tokens, with a reasonnable computational cost.

Overall, several links can be made with other work. Especially, the work by Kassner et al. (2021) is the closest to ours and also uses distractors to assess factuality in LMs. However, their set of distractors for assessment is limited to a little more than 30,000. In practice, though, there are millions of entities in Wikidata, highlighting the need for effective retrieval strategies to avoid performing inference on each distractor. Then, the idea to rely on distractor is close to the task of multiple-choice question answering (MCQA) (Hendrycks et al., 2021; Boratko et al., 2018) since distractors can be seen as incorrect answers. MCQA assumes an instruction-tuned model and introduces the ambiguity that errors may stem either from the model's inability to handle questions or from a deeper lack of knowledge. Still, the shared issue is to automatically determine the possible responses that will most challenge the LM.

Finally, let one highlight that temporality of facts is an important dimension that is under-explored in the knowledge assessment literature. However, we postpone this aspect for future work in order to focus, yet, on measuring factual knowledge valid at the present time, which is a simpler but still challenging problem.

3 Our Knowledge Measure

The key principle of our knowledge measure is that a model "knows" a fact (s, r, o) if it prefers (s, r, o) against plausible incorrect alternatives of the form (s, r, o^*), where the objects o^* are referred to as

Know. measure	Easily interpretable	Robust to OOS	Comparable inter-LM	Support multi-token answers	Computation cost
ROUGE-L	✓	✗	✓	✓	++
BERT-score	✗	✗	✓	✓	++
LLM-as-a-judge	✓	✗	✓	✓	+++
Precision@n	✓	✗	✗	✗	+
Probability	✗	✓	✗	✓	+
KaRR	✗	✓	✗	✓	++
Distractors (ours)	✓	✓	✓	✓	++

Table 1: Pros and cons of each knowledge measure.

distractors. Evaluating the preference relies on verbalizing the pair (s, r) as a cloze sentence and measuring the likelihood of the objects o and o^* as possible continuations.

This section first formalizes the knowledge measure and its key concepts. Then, it presents various strategies to retrieve distractors in anticipation of experiments in Section 4.1.

3.1 Fact Verbalization

Verbalization consists in mapping facts into natural language. This is a key step since facts are symbolic objects in knowledge bases and LMs can only handle textual data . For our purposes, we decompose the verbalization of a fact (s, r, o) into two aspects: the pair (s, r) can be mapped to a cloze sentence; the entity o can be expressed in one or several ways. For both aspects, the main problem is to handle the variability of natural language as it has been shown that LMs can sometimes answer differently to semantically equivalent variants of the same prompt (Elazar et al., 2021; Kassner and Schütze, 2020). Hence, we assume that a given pair (s, r) can be associated with a set of semantically equivalent cloze sentences, noted as $C(s, r)$. Likewise, an entity e can be expressed in different forms, or *labels* in the terminology of knowledge bases (for instance, a person being referred to through her/his full name or only last name). Thus, an entity e is associated with a set of labels $\Lambda(e)$.

3.2 Knowledge Measure $\mathcal{K}$

Our knowledge measure is defined as a function $\mathcal{K}$ that takes as input a fact $f = (s, r, o)$. For simplicity, let us consider a single cloze sentence c of $C(s, r)$ to assess knowledge of (s, r, o), our measure relies on the probability to continue c with a label of o. The sum of these probabilities is referred to as the *plausibility* $\mathrm{Pl}(o|c)$ and is defined as:

$$\mathrm{Pl}(o \mid c) = \sum_{\lambda \in \Lambda(o)} \Pr(\lambda + EOS \mid c) \qquad (1)$$

where EOS is the end of sentence token.

The plausibility of o is then compared to the plausibility of several *distractors*. A distractor for o is defined as an entity o^* which has the same type as o, no common labels with o^*, and such that (s, r, o^*) is incorrect. Given a set $\Delta_n(f)$ of n distractors for the fact f, if the model knows f, then we should observe $\mathrm{Pl}(o|c) > \mathrm{Pl}(o^*|c)$ for any distractor $o^* \in \Delta_n(f)$. Thus, we define the measure $\mathcal{K}$ as follows:

$$\mathcal{K}(f|c) = \underset{o^* \in \Delta_n(f)}{\mathrm{Agg}} \mathbb{1}\left[\mathrm{Pl}(o|c) > \mathrm{Pl}(o^*|c)\right] \quad (2)$$

where $\mathbb{1}$ denotes the indicator function, and Agg is an aggregation function.

Choosing the *minimum* as an aggregation function is a natural choice, which yields a strict measure of knowledge, where the score is 1 if o is preferred to all distractors, 0 otherwise. Another reasonable choice is the *average*, which can be seen as a smooth version of *minimum*. We refer to the knowledge measures using these two aggregation functions as **Min@n** and **Avg@n** respectively.

Finally, generalizing to all cloze sentences $C(s, r)$ the final formula for $\mathcal{K}$ is defined as follows:

$$\mathcal{K}(f) = \frac{1}{|C|} \sum_{c \in C} \mathcal{K}(f|c) \, . \qquad (3)$$

3.3 Distractor Retrieval Strategies

We consider a knowledge base $\mathcal{B}$ defined as follows. Let E be set of entities and R a set of relations. Then $\mathcal{B} \subseteq E \times R \times E$, i.e., $\mathcal{B}$ is a set of triples of the form (entity, relation, entity). Moreover, we consider a function types that maps each entity to its types, e.g., $\text{types}(Paris) = \{City\}$. For simplicity, we assume that $\mathcal{B}$ contains all true facts of $E \times R \times E$; in other words: a fact f is correct if and only if $f \in \mathcal{B}$.

The set of all distractors of f can be formalized as the set $\Omega(f)$ of all entities $o^* \in E$ such that:

- $(s, r, o^*) \notin \mathcal{B}$;
- $\text{types}(o^*) \cap \text{types}(o) \neq \varnothing$; [†]
- $\Lambda(o) \cap \Lambda(o^*) = \varnothing$.

For any input f, the value of $\mathcal{K}(f)$ depends on the set of n considered distractors $\Delta_n(f)$. Here, Δ denotes a *distractor retriever*, i.e., a function that maps any input fact f to a subset of $\Omega(f)$.

The retrieval strategies we considered are described in what follows:

Optimal Distractors. Given the objective of the distractors to be competitive alternatives for a given LM and a triple (s, r, o), the best possible distractors are the entities of the same type as o and with the highest conditional probability given (s, r) as provided by the LM. We denote the set of the n most plausible distractors as:

$$\Delta_n(f) = \underset{o^* \in \Omega(f)}{\textbf{top-n}} \ \text{Pl}(o^*|c) \ . \tag{4}$$

Finding an exact solution to this problem involves exploring a huge portion of the generation tree, which can be very expensive.

Approximation of Optimal Distractors. This strategy (noted **ApprOpt**) gives an approximation of the set of optimal distractors defined in Equation 4. It uses a beam search of width equal to n as a decoding strategy, constrained using a grammar (Geng et al., 2023; Cao et al., 2021) enforcing the generation of a continuation of the form λ + EOS, where λ is a label of o or of one of its distractors. It has to be noted that the number of retrieved distractors can be smaller than n because it is possible for a label of o to be encoded in two different sequences of tokens. A variant of grammar-constrained decoding that avoids this problem is an interesting direction for future work.

Semantic Distractors. Searching for distractors in the LM generation tree is costly; a more frugal approach is to frame distractor retrieval as a semantic search task. The idea is to select distractors that share common properties with o, by adapting the classical TF-IDF approach[‡]. In a first step, all entities $e \in E$ are encoded as a bag of features containing: (i) the entity e itself, (ii) all r such that $(e, r, o) \in \mathcal{B}$ for some o; (iii) all o such that $(e, r, o) \in \mathcal{B}$ for some r; (iv) all (r, o) such that $(s, r, o) \in \mathcal{B}$. Then, all bags of features are represented as TF-IDF vectors[§] (Salton and Buckley, 1988), by considering each entity in E as a document. At retrieval time, the similarity between two entities is computed as the cosine similarity between their respective vectors. For a given fact $f = (s, r, o)$, the method returns the n most similar distractors in $\Omega(f)$. This strategy is noted **Sem**.

Temporal+Semantic Distractors. Temporal distractors are objects o^* such that (s, r, o^*) was valid in the past but is not anymore at the present time. For example, the temporal distractors of the pair $(\textit{USA}, \textit{president})$ are all the presidents of USA except the current one. Since the training data of LMs spans over multiples years/decades, verbalizations of facts with temporal distractors have probably been observed during the LM's pretraining. Temporal distractors also benefit from being most likely semantically connected to o. In this strategy (noted **Temp+Sem**), temporal distractors are prepended to the list of semantic distractors. We note that *Temp+Sem* and *Sem* strategies have the advantage over *ApprOpt* of producing the same distractors for all LMs, making their performance comparable.

Random Distractors. This last strategy draws random entities $o^* \in \mathcal{B}$ of a common type with o, still under the constraint $(s, r, o^*) \notin \mathcal{B}$. Random distractors are considered to provide a lower bound on the quality of the distractors.

The next section aims at comparing between the different retrieval strategies.

4 Comparison of Knowledge Measures

This section first compares retrieval strategies to determine which one produces the hardest distrac-

[†] For simplicity, the types of an entity are restricted to those that were explicitly declared in Wikidata via the relations *instance of* and *subclass of*, i.e., $\text{types}(e)$ do not contain the whole type hierarchy of e. Otherwise, all entities would have the Entity type in common.

[‡] TF-IDF is not essential for this strategy to work, it is only a representative of the family of weighing methods and is used to evaluate semantic search methods in general. Thus, another weighing technique, such as BM25 (Robertson and Zaragoza, 2009), can be used instead.

[§] The TF-IDF vectors are indexed using NMSLIB for fast retrieval (`github.com/nmslib/nmslib`).

tors for LMs. It then compares the strategies with measures from the literature based on their correlation with human knowledge assessment. Finally, comparisons are also presented regarding the robustness to verbalization artifacts, as verbalization often rely on templates.

4.1 Which strategy produces the hardest distractors?

In this experiment, our knowledge measure (Min@n) is applied with several distractor retrieval strategies to assess Pythia-6.9B's (Biderman et al., 2023) knowledge of a set of triples. Since good retrieval strategies should yield challenging distractors, the best retrievers are those that produce the lowest Min@n values.

Data. The experiments rely on a preprocessed version of the 2021-01-04 dump of Wikidata (Ammar Khodja et al., 2024), which provides triples with their popularity scores, as well as the types and labels of each entity involved. It features 51 million triples, 2,100 relations, 10 million unique entities, and 1.34 labels per entity on average, indicating a large coverage of factual knowledge. Moreover, each relation comes with many templates for verbalizing facts in natural language. These templates were generated using post-processed GPT3.5 fact verbalizations. We augment the number of templates to increase the coverage of Wikidata relations from 1,123 to 1,866 relations, which is the richest template database to our knowledge.

A set S of 1000 facts is sampled with various levels of popularity. For each fact $(s, r, o) \in S$, a cloze sentence is generated for (s, r) by filling the subject slot of the template of r with a label of s.

Results. Figure 1 displays the mean value of Min@n over S for n varying from 1 to 100. We first note that the distractors retrieved using *Temp+Sem* are significantly harder than random entities from the target type (*Random*). The *ApprOpt* strategy produces the most challenging distractors, in particular when n is small. Indeed, the first distractor deceives the LM (Pythia-6.9B) in 60% of the time on average, highlighting the poor factual accuracy of LMs even for this kind of model size. As n increases, Min@n does not converge to zero but stabilizes at 0.35, meaning that there is a portion of facts that is robust to **ApprOpt** distractors. The curve of the *Sem* was not plotted because it was identical to the *Temp+Sem* curve, which was initially surprising. After further analysis, we found that only $\sim$ 1% of facts in Wikidata possess at least one temporal distractors, leading to no difference in practice between these two strategies.

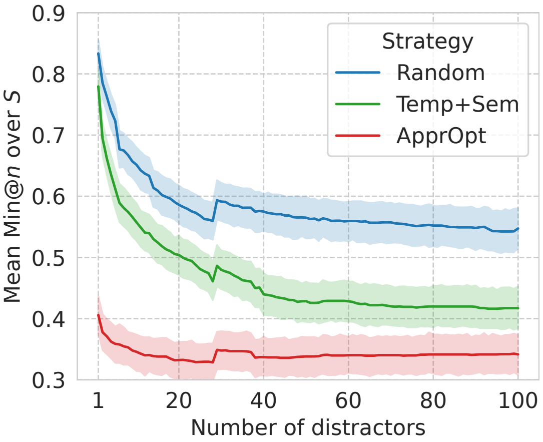

Figure 1: Comparison of the different retrieval strategies on S using Pythia-6.9B (the lower the curve, the harder the produced distractors for the LM).

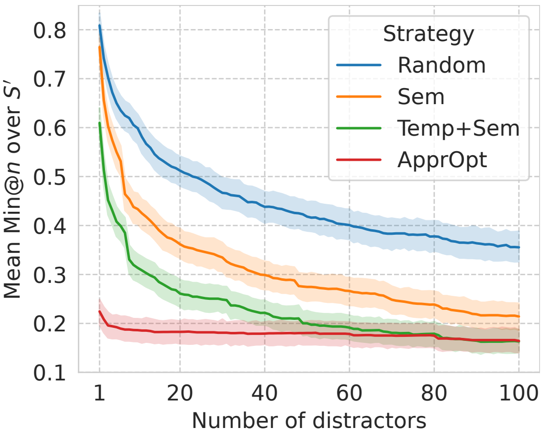

Figure 2: Comparison between the difficulty of the retrieval strategies on S' using Pythia-6.9B.

To better assess the advantage of temporal distractors, A set of 1000 facts, noted S' from Wikidata is sampled following the same procedure to sample S, except that the kept facts possess at least 1 temporal distractors. On S', *Temp+Sem* produced distractors that are significantly harder than *Sem*'s, indicated by a decrease of 0.17 in Min@1 (Figure 2). We conclude that *Temp+Sem* is a superior strategy compared to *Sem*, producing distractors at least as difficult as those of *Sem*.

4.2 How does it compare with other knowledge measures?

In the following two sections, our knowledge measure is compared with several measures found in the literature regarding two aspects: their correlation with human judgement, and their sensitivity to verbalization errors.

Metrics. As already introduced in Section 2, the studied measures are : Probability, Precision@n (Petroni et al., 2019), BERT-score (Zhang et al., 2020), ROUGE-L (Lin, 2004), KaRR (Dong et al., 2023), and LLM-as-a-judge (Sun et al., 2024). Their formal definition is provided in Table 2 with the same formalism as our method. LLM-as-a-judge should be distinguished from other methods because it is far more expensive, since one objective of our work is to provide a metric that can be computed using a reasonable amount of resources. The complexity of KaRR's equation prevented its inclusion in Table 2. For details about this measure, readers should refer to the original paper Dong et al. (2023).

Here, the implementation of LLM-as-a-judge is based on the prompt in Sun et al. (2024) (in Appendix A.1.2 of this reference). This prompt was adapted to our data since we use declarative cloze sentences for this experiment (Appendix D), and we choose the judge to be GPT3.5[¶].

Data. Dong et al. (2023) provides a dataset where humans assessed the knowledge of GPT2-XL (Radford et al., 2019) on 210 facts[‖]. Two groups of annotators were involved. For each fact (s, r, o), the first group was charged to prompt GPT2-XL using cloze sentences until the response of the model was of same type as o, while the second group rated the response (0 if incorrect and 1 if correct). The human assessment score is the average rating across all human scores from the second group, and all prompts produced by the first group, which ranges from 0 to 1. The reported Kappa score of the annotations of the second group is 0.4. For a fair comparison, all the measures use the same cloze sentences, 5 per fact, to assess knowledge.

Experiment. Because several of these measures have hyperparameters, a phase of calibration is performed in order to maximize the correlation to the human knowledge assessment scores. Following previous work (Dong et al., 2023), the correlation measure that is used is Kendall's τ. The alignment with human judgement of each knowledge measure is computed as the mean Kendall's τ across the $K = 3$ folds of cross-validation.

This optimization process is done using a grid search over all hyperparameters for each measure: the beam width parameter is varied in {1, 2, 5, 10, 20, 30, 50, 100} except in LLM-as-a-judge, where it is varied in {1, 2, 5}, the number of distractors is varied in {1, 2, 5, 10} ∪ {20, 40, ..., 200}, as well as n in Precision@n.

Results. The correlation to human judgment for each knowledge measure is shown in Table 3 and the best configurations per fold are shown in the Appendix. We were unable to reproduce the reported KaRR performance of 0.42 in Kendall's τ with human judgement in our experiments and using our facts. We report however our tentative in Table 3 which is much lower.

First, methods that explore the generation tree (ROUGE-L and BERT-score) are not well correlated to human judgement, except for LLM-as-a-judge which is surprisingly well correlated given that it used a beam width of only 1, 2, and 2 respectively for each fold. Its main inconvenient remains the cost of this approach compared to other measures. As of our approach, the best strategy retrieval is *ApprOpt* followed by *Temp+Sem*, then finally *Random*. We note also that *Probability*, which consists simply taking the average of probability of the correct objects, correlates surprisingly well with human judgement compared to other metrics.

The best configurations of each knowledge measure per test fold are kept for the next experiments.

4.3 Robustness to verbalization artifacts

A neglected aspect in the literature is the importance of the verbalization process in the quality of the assessment. In general, the use of templates introduce several problems, because syntactical artifacts can arise, such as missing or mispresence of determiners or prepositions before the object. For instance, the template *"[SUB] is located in the [OBJ]"* works for the triple (Washington, location, United States) but fails for the triple (Tokyo, location, Japan) because *"the Japan"* is syntactically incorrect. This section evaluates the robustness of knowledge measures to these errors by assessing how much they deviate between using a perfect verbalization compared to a flawed one. Given a knowledge measure, a small deviation indicates that template-based verbalization can effectively substitute a manually crafted high-quality verbalization, while a high deviation indicates that this measure is too sensitive to verbalization artifacts.

[¶] *gpt-35-turbo-16k-0613*

[‖] This is a subset of the full dataset of 410 assessed facts that was used in their own work.

Name	Definition	Hyperparameters
Our measure	$\frac{1}{\lvert C\rvert}\sum_{c\in C} \underset{o^*\in\Delta_n(f)}{\text{Agg}} \mathbb{1}\left[\text{Pl}(o\lvert c) \ > \ \text{Pl}(o^*\lvert c)\right]$	Aggregation function, number of distractors
Precision@n	$\frac{1}{\lvert C\rvert}\sum_{c\in C} \mathbb{1}\left[\left(\underset{\gamma\in V}{\textbf{top-k}}\Pr(\gamma \mid c)\right)\cap\Lambda(o)\neq\varnothing\right]$	n
Probability	$\frac{1}{\lvert C\rvert}\sum_{c\in C}\sum_{\lambda\in\Lambda(o)}\Pr(\lambda\lvert c)$	-
BERT-score	$\frac{1}{\lvert C\rvert}\sum_{c\in C}\max_{\lambda\in\Lambda(o),b\in B(c)}\text{BERTScore}(\lambda,b)$	Beam width
ROUGE-L	$\frac{1}{\lvert C\rvert}\sum_{c\in C}\max_{\lambda\in\Lambda(o),b\in B(c)}\text{ROUGE}_L(\lambda,b)$	Beam width
LLM-as-a-judge	$\frac{1}{\lvert C\rvert}\sum_{c\in C}\max_{b\in B(c)}\text{LLMJudge}(\Lambda(o),b)$	Beam width

Table 2: Knowledge measures definitions. C is the set of cloze sentences used to assess the fact $f=(s,r,o)$. $B(c)$ are the best sequences obtained from the LM using beam search given the prompt c. V is the vocabulary of the LM. The complexity of KaRR's equation prevented its inclusion in this table.

Know. measure	Mean τ	Std τ
KaRR†	0.104	0.039
ROUGE-L	0.138	0.122
BERT-score	0.159	0.063
Precision@n	0.185	0.028
Probability	0.262	0.047
Dist. Random	0.225	0.021
Dist. Temp+Sem	0.242	0.023
Dist. ApprOpt	**0.282**	0.024
LLM-as-a-judge	0.293	0.029

Table 3: Correlation of knowledge measures to human judgement. The measure with the maximum correlation (excluding LLM-as-a-judge) is in **bold**. † our tentative of measuring the correlation of KaRR. The metrics were grouped as : baseline knowledge measures (top group), our measures (middle group), and expensive baseline measures (bottom group)

Data. To build a dataset of verbalization errors, we sampled 77 facts from S (Section 4.1) and verbalized each one of them using one of the five available templates. An annotator (one of the authors) then reviewed each verbalization, identified any errors, and proposed corrected versions, while being as critical as possible. The results showed that 45.5% of the verbalizations contained at least one minor error (Wilson's 95% CI = [34.8, 56.5]), **indicating that small artifacts are very common in template-based verbalization**. However, no major errors were reported (Appendix B).

Know. Measure	Robustness
ROUGE-L	0.47
LLM-as-a-judge	0.58
BERT-Score	0.61
Probability	0.63
Dist. ApprOpt	**0.92**

Table 4: Robustness of knowledge measures to verbalization artifacts.

Experiment. To measure robustness to these artifacts, we kept the 32 flawed verbalizations along with their correct counterparts and fed them to the studied measures. The robustness of a knowledge measure to verbalization artifacts is determined by how similar the outputs are between the correct and incorrect verbalizations. This similarity is computed using Kendall's τ averaged over the best configurations of each measure (Table 4).

Results. Although the *Probability* measure is almost as correlated to human judgement as our measure, it is more sensitive to verbalization artifacts. Conversely, *Dist. ApprOpt* is the by far the least sensitive with a Kendall's τ of 0.92, making our measure the one that is well correlated to human judgement, while being the most robust to artifacts.

5 Exploration

In this section, we delve deeper into the factors influencing the alignment of our studied measures

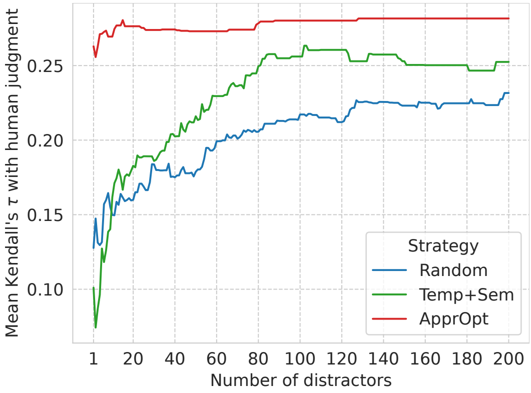

Figure 3: Relation between the number of distractors and the correlation to human judgement.

with human judgment, in a first part, before studying how model size influences the robustness to distractors in a second part.

5.1 Influence of different factors

We study the impact of the number of distractors for our knowledge measure, and the impact of the number of cloze sentences for all measures. This is useful in resource-limited applications to determine how much correlation is lost when reducing either distractors or cloze sentences.

Impact of the number of distractors. We group the best configurations (from Section 4) of our knowledge measure per retrieval strategy, and we vary the number of distractors from 1 to 200, while computing the mean correlation with human judgement for each number over the cross-validation folds. The results appear in Figure 3.

The *Random* and *Temp+Sem* retrieval strategies reach optimum at around 130 and 110 distractors respectively. On the other hand, *ApprOpt* strategy requires less and converges rapidly to a near-optimal correlation at approximately 15 distractors.

Impact of the number of cloze sentences. We group the best configurations of the studied knowledge measures (from Section 4) and we vary the number of cloze sentences from 1 to 5, while computing the mean correlation with human judgement for each number over the cross-validation folds. Results are shown in Figure 4.

All knowledge measures benefit from a larger number of cloze sentences, indicating the importance of varying verbalizations to obtain a score representative of human rating. This is especially the case for BERT-score, ROUGE-L, and LLM-as-a-judge which proportionally gain +5%, +3%, and +7%, of correlation respectively. It appears that *ApprOpt* reaches its peak at 3 sentences, indicating that this number is sufficient for an optimal correlation with human judgement.

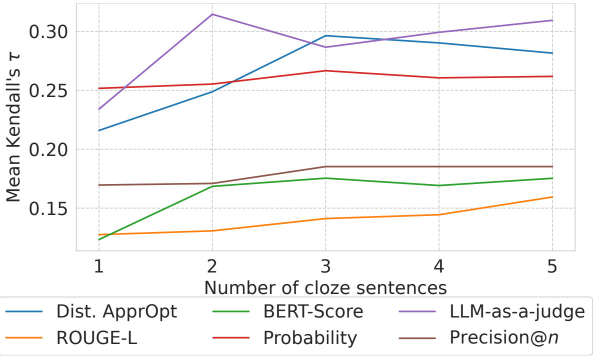

Figure 4: Relation between the number of cloze sentences and the correlation to human judgement.

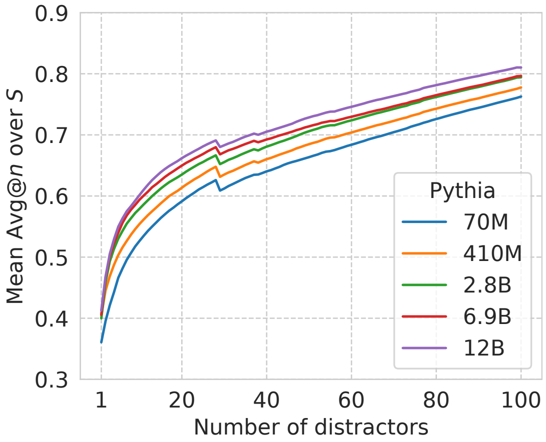

Figure 5: Mean knowledge measure over S on different model sizes using the **ApprOpt** retrieval strategy

5.2 Are larger models more robust to distractors?

In this section, we use our measure to investigate the robustness of language models to *ApprOpt* distractors and how it scales with the LM size.

To achieve this, we repeat the experiment from the Section 4, varying the model size within the Pythia family and using *ApprOpt* as the retrieval strategy. Pythia models are ideal because while their size changes, they are trained on the same number of tokens and iterations.

In Figure 5, a trend emerges: the larger the model, the more robust it is to distractors. However, the gain in performance as the model size grows is quite low. LMs are inherently vulnerable to distractors and increasing the model size helps but is insufficient to attain a satisfactory robustness.

6 Conclusion

In this paper, we introduced a new interpretable measure to assess factual knowledge in language models, by leveraging distractors. We defined and compared different retrieval strategies in terms of the difficulty of the distractors that find. We compared our measure with others in the field on the base of theoretical arguments (Table 1), human judgment correlation and robustness to verbalization artifacts. Our results demonstrate that our distractor-based knowledge measure aligns well with human knowledge assessment, while being the most resilient to verbalization artifacts. Overall, our distractor-based knowledge measure offers a promising direction for more accurately evaluating the factual knowledge embedded in LMs.

7 Limitations

Human Annotations. Because human validations are expensive to obtain, we rely on a limited set of human annotations to validate our conclusions. A larger human knowledge assessment dataset and verbalization errors dataset are advisable to solidify our findings.

Out-of-subject continuations can be informative. Out-of-subject continuations, that are generated by the LM, were assumed to be uninformative of whether they know the tested fact. While in fact, they can be informative and at different degrees. For example, assuming an LM is tested on the fact (Paris, capital of, France), the LM continuing "Paris is the capital of" with "a country", is less informative than "a European country", which is less informative than "France". We did not explore this aspect as it is very difficult to measure precisely the information provided by a continuation.

8 Ethical Considerations

The detection of misinformation and inaccuracies in language models is a significant challenge in artificial intelligence research. Our work contributes to this goal by offering a tool and resources designed to assess the factuality of LMs using reliable information from the Wikidata knowledge base, and could potentially contribute to the enhancement of the truthfulness and reliability of these models. This is especially crucial in an era where misinformation can spread rapidly through digital channels.

However, it should be remembered that while Wikidata is a reliable source, it is not immune to biases and inaccuracies that may arise from the contributions of its vast user base.

References

Hichem Ammar Khodja, Frederic Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2024. WikiFactDiff: A large, realistic, and temporally adaptable dataset for atomic factual knowledge update in causal language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17614–17624, Torino, Italia. ELRA and ICCL.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 60–70. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguistics*, 9:1012–1031.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2023, Singapore, December 6-10, 2023, pages 10932–10952. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. 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.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7811–7818. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Kai Sun, Yifan Ethan 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 311–325. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

A Extending the template database of WikiFactDiff

WikiFactDiff (Ammar Khodja et al., 2024) is a dataset that describes the evolution of factual knowledge between 2021 and 2023, where facts are represented as triples. This dataset also provides a database of templates to convert these facts to natural language, covering 1123 Wikidata relations. To build this database, the authors asked GPT3.5 to verbalize 26K facts from Wikidata, which were then post-processed to produce many templates for each relation.

To cover a larger number of relations, we increase the number of verbalized facts to 437954 (346702 from triples in WikiFactDiff and 91252 from Wikidata) using the same procedure as in WikiFactDiff, resulting in a coverage of 1866 relations.

B Taxonomy of Verbalization Errors

This section describes in detail the annotation process of verbalization errors and their correction, to assess their impact on the studied knowledge measures (Section 4.3).

The annotation process involves an NLP expert (an author of this paper) and it is performed as follows. First, 77 triples with various levels of popularity are sampled from the preprocessed Wikidata dump of 2021-01-04 and verbalized using a random template. Then, the annotator iterates through each verbalization with access to the triple being verbalized. Then, for each verbalization, the annotator detects the errors, defines them in Table 8, and corrects the verbalization. It has to be noted that the annotator was asked to be as critical as possible when annotating errors. Each verbalization can have zero, one, or many errors at the same time.

Now that each verbalization has an annotation and potentially a correction, the proportion of each error are computed and reported in Table 5.

To produce the summary shown in Table 6, we group all the errors in three categories:

- **minor errors** : obj_deter, sub_deter, obj_unclear, sub_unclear, adj_noun_conjug, better_verb, wrong_conjug
- **blunders** : out_of_subject, wrong_language

We note that most errors concern the bad usage of determiners which are probably the easiest errors to correct automatically in this list.

Error type	Proportion (%)	Wilson's CI
adj_noun_conjug	1.3	(0.2, 7.0)
better_verb	3.9	(1.3, 10.8)
obj_deter	13.0	(7.2, 22.3)
obj_unclear	5.2	(2.0, 12.6)
out_of_subject	0.0	(-0.0, 4.8)
sub_deter	14.3	(8.2, 23.8)
sub_unclear	5.2	(2.0, 12.6)
wrong_conjug	14.3	(8.2, 23.8)
wrong_language	0.0	(-0.0, 4.8)
Free of errors	**54.5**	**(43.5, 65.2)**

Table 5: Proportion (%) of the presence of each error in the verbalizations produced by our verbalizer with their respective 95% Wilson's confidence interval

Error type	**Proportion (%)**	**Wilson's CI**
only minor errors	45.5	(34.8, 56.5)
contain blunder(s)	0.0	(0.0, 2.5)
Free of errors	**54.5**	**(43.5, 65.2)**

Table 6: Proportion (%) of error categories in the verbalizations produced by our verbalizer with their respective 95% Wilson's confidence interval

C Which measure benefits the most from a large number of cloze sentences?

Proportional gain in Kendall's τ for a particular knowledge measure is defined as $\frac{\tau_m + 1}{\tau_1 + 1}$, where τ_m is the mean Kendall's τ with human judgement over the test folds of the cross validation, using m cloze sentences. We plot the proportional gain in Figure 6. We conclude from the results that decoding methods (BERT-score, ROUGE-L, and LLM-as-a-judge) are the ones that benefit the most from multiple cloze sentences which can be explained by the fact that a larger number of cloze sentences reduces the probability of falling in an OOS continuation when decoding, giving more information to the knowledge measure to judge the LM's knowledge. On the other hand, our knowledge measure, which is well correlated to human knowledge, reaches the peak of correlation after 3 cloze sentences, which indicates that our method is frugal with respect to the needed number of close sentences.

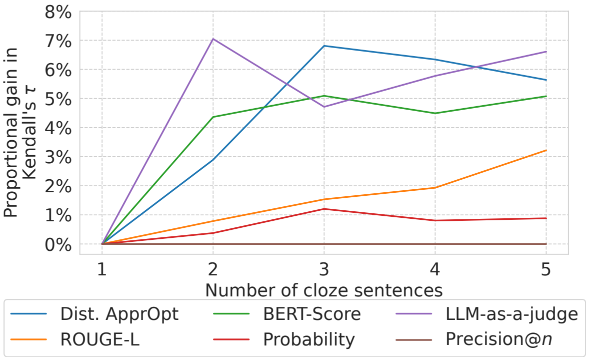

Figure 6: Proportional gain in Kendall's τ with respect to the number of cloze sentences.

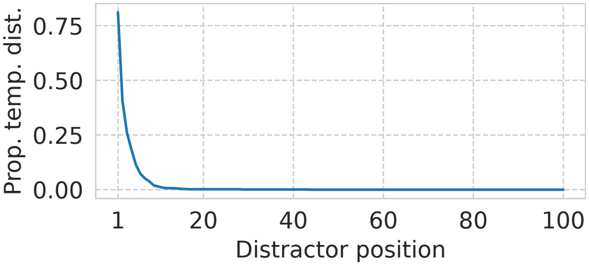

Figure 7: Proportion of temporal distractors per distractor position in S'.

D LLM-as-a-judge prompt

We adapt the prompt in Sun et al. (2024) which used questions to assess the LM's knowledge to fit our data which contains cloze declarative sentences.

You need to check whether the prediction of a question-answering system to a query is correct. You should make the judgment based on a list of ground truth answers provided to you in a form of a list of aliases of the gold answer. Your response should be "correct" if the prediction is correct or "incorrect" if the prediction is wrong.

Query: The author of The Taming of the Shrew (published in 2002) is ____
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: W Shakespeare
Correctness: correct

Query: The author of The Taming of the Shrew (published in 2002) is ____
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: Roma Gill and W Shakespeare
Correctness: correct

Query: The author of The Taming of the Shrew (published in 2002) is ____
Ground truth: ["William Shakespeare", "Roma Gill"]
Prediction: Roma Shakespeare
Correctness: incorrect

Query: The country where Maharashtra Metro Rail Corporation Limited is located is ____
Ground truth: ["India"]
Prediction: Maharashtra
Correctness: incorrect

Query: The job of Song Kang-ho in Parasite (2019) is ____
Ground truth: ["actor"]
Prediction: He plays the role of Kim Ki-taek, the patriarch of the Kim family.
Correctness: correct

Query: The era to which Michael Oakeshott belongs is ____
Ground truth: ["20th-century philosophy"]
Prediction: 20th century.
Correctness: correct

Query: The department where Edward Tise (known for Full Metal Jacket (1987)) was is ____
Ground truth: ["sound department"]
Prediction: 2nd Infantry Division, United States Army
Correctness: incorrect

Query: The wine region to which Finger Lakes AVA belongs is ____
Ground truth: ["New York wine"]
Prediction: Finger Lakes AVA
Correctness: incorrect

Query: [QUERY]
Ground truth: [LIST OF REFERENCES OF CORRECT OBJECT]
Prediction: [LM PREDICTION]
Correctness:

E Extrapolating the relation between LM size and robustness to distractors

In Figure 5, a trend emerges: the larger the model, the more robust it is to distractors. However, the gain in performance as the model size grows is quite low. Extrapolating this evolution with a log-scale linear regression that fit very well the data ($R^2 = 0.995$), an absurd number of 2×10^{19} (twenty quintillion) parameters is needed to achieve a 90% Avg@20 (Figure 8). This indicates that reasonably increasing the size of LMs is not a solution to this problem.

Random Distractors	**Semantic Distractors**	**Temporal+Semantic Distractors**	**Approx. of Optimal Distractors**
Hugo Heermann	Barron Trump	Barack Obama	Barack Obama
Suzanna von Nathusius	Donald Trump Jr.	George W. Bush	George W. Bush
Bernhard Heinrich Overberg	Ivanka Trump	Bill Clinton	John F. Kennedy
Joseph Siffert	Eric Trump	George H. W. Bush	William Howard Taft IV
Bep Vriend	Mary Anne MacLeod Trump	Ronald Reagan	Kim Jong-un
Lucien Carr	Tiffany Trump	Jimmy Carter	Donald Trump Jr.
Andries van Dam	Joe Biden	Gerald Ford	George H. W. Bush
Gilles Simeoni	Melania Trump	Richard Nixon	Michael Phelps
Ghatam Udupa	Marla Maples	Lyndon B. Johnson	Donald Rumsfeld
Kris Long	Haim Saban	John F. Kennedy	Mike Ditka
Maxlei dos Santos Luzia	Mark Warner	Dwight D. Eisenhower	Colin Kaepernick
Yael Lotan	Michael Eisner	Harry S. Truman	John F. Kennedy Jr.
Dorman Bridgman Eaton	Barack Obama	Franklin Delano Roosevelt	Thomas S. Monson
Harry Biedermann	George Lindemann	Herbert Hoover	Dmitry Medvedev
Ischke Senekal	Morgan Fairchild	Calvin Coolidge	Mikhail Gorbachev

Table 7: Sample of distractors retrieved using different retrieval strategies given the fact (USA, president, Donald Trump). The cloze sentence used for the ApprOpt strategy is *The president of USA is* ____ and the model used is Pythia-12B.

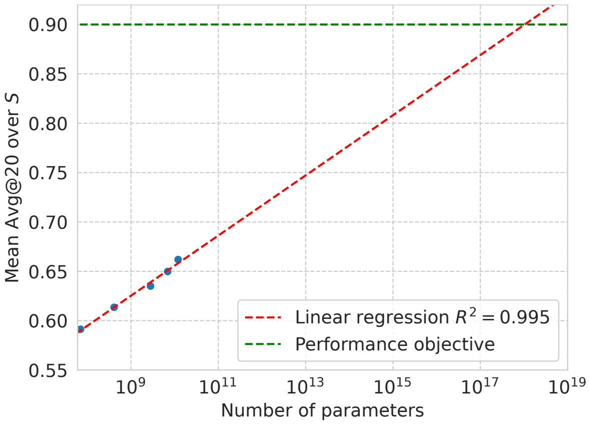

Figure 8: Relation of model size to Avg@20 using the **ApprOpt** retrieval strategy

Name	**Description**	**Example**
adj_noun_conjug	A noun or adjective is mistakenly considered to be the verb to conjugate	The language **uses** in France is French
better_verb	The verbalization does not perfectly conveys the fact in the triple and could be mistaken for another meaning	The language used in Nigeria is Koenoem *(except that Nigeria has 525 native languages and Koenoem is not the predominant one)*
(sub|obj)_deter	There is a determiner missing before the subject or object	Alexander Macomb House was where *(the)* president of the United States officially resides
(sub|obj)_unclear	The subject or object cannot be easily determined and “blends” with the sentence because it is too long for instance	Nikhil Dwivedi is part of the cast of Scam 1992: The Harshad Metha story
wrong_conjug	The verb is conjugated in the wrong tense.	Last year, the president of USA **will be** Joe Biden or Bluetooth **is** created in Sweden *(“was” is better)*
wrong_language	The verbalization is performed in a language other than English	Il presidente degli Stati Uniti è Joe Biden
out_of_subject	The verbalization is out-of-subject and does not convey the fact in the triple at all	-

Table 8: Definition of the verbalization errors

Know. measure	Retr. strategy	Agg.	Num. distractors	Beam width	n	τ
BERT-score				100		0.21
				1		0.09
				2		0.17
Dist.	ApprOpt	Min@n	200			0.31
						0.26
						0.27
	Sem	Min@n	80			0.22
			200			0.27
			120			0.24
	Random	Min@n	200			0.24
						0.20
			180			0.23
	Temp+Sem	Min@n	80			0.22
			200			0.27
			120			0.24
LLM-as-a-judge				1		0.27
				2		0.28
						0.33
Probability						0.32
						0.24
						0.23
Precision@n					100	0.17
						0.17
						0.22
ROUGE-L				100		0.25
				1		0.01
						0.15

Table 9: Best configuration of each knowledge measure per cross-validation fold with respect to the Kendall’s τ correlation with human judgement.