Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators

Matéo Mahaut*
Universitat Pompeu Fabra
mateo.mahaut@upf.edu

Laura Aina **Paula Czarnowska** **Momchil Hardalov**
Thomas Müller **Lluís Màrquez**
AWS AI Labs
{eailaura, czarpaul, momchilh, thomul, lluismv}@amazon.com

Abstract

Large Language Models (LLMs) tend to be unreliable in the factuality of their answers. To address this problem, NLP researchers have proposed a range of techniques to estimate LLM's confidence over facts. However, due to the lack of a systematic comparison, it is not clear how the different methods compare to one another. To fill this gap, we present a survey and empirical comparison of estimators of factual confidence. We define an experimental framework allowing for fair comparison, covering both fact-verification and question answering. Our experiments across a series of LLMs indicate that trained hidden-state probes provide the most reliable confidence estimates, albeit at the expense of requiring access to weights and training data. We also conduct a deeper assessment of factual confidence by measuring the consistency of model behavior under meaning-preserving variations in the input. We find that the confidence of LLMs is often unstable across semantically equivalent inputs, suggesting that there is much room for improvement of the stability of models' parametric knowledge. Our code is available at https://github.com/amazon-science/factual-confidence-of-llms.

1 Introduction

A major problem of Large Language Models (LLMs) is that they do not always generate truthful information. Models can hallucinate by convincingly reporting information that is actually false or they are not confident about, or provide factual answers only when prompted in a certain way (Elazar et al., 2021; Lin et al., 2022b; Ji et al., 2023; Wang et al., 2023a; Luo et al., 2023). This behavior can be severely harmful, especially given the current explosion of LLM usage: a lack of truthfulness can lead to spread of misinformation and breaches to user trust (Weidinger et al., 2021; Bender et al., 2021; Evans et al., 2021; Tamkin et al., 2021). Having a reliable estimate of the model's confidence over a fact—the degree to which it is expected to have accurate factual knowledge with respect to an input—is key for mitigating this problem (Geng et al., 2023; Tonmoy et al., 2024).

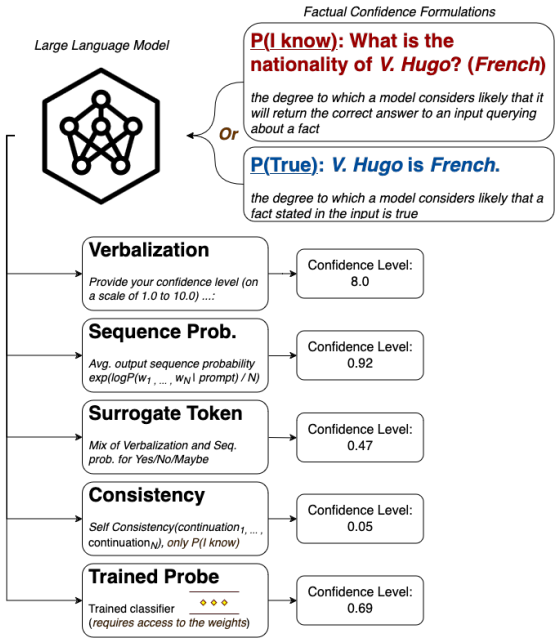

Figure 1: Overview of our factual confidence estimation framework. We work with five groups of methods and two formulations: $P(\text{I know})$, which applies to questions, and $P(\text{True})$, which applies to statements. All of the methods produce a continuous score, except *verbalization*, where the model generates a confidence level.

Recently, a number of methodologies for estimating the factual confidence of LLMs were proposed (Lin et al. 2022a; Burns et al. 2023; Kuhn et al. 2023; Azaria and Mitchell 2023; Pacchiardi et al. 2024, among others). However, none of them establishes a unified experimental framework to compare methods. This leaves open questions regarding how aligned the methods are in their estimates, and which are the most reliable ones across models. We aim to fill this gap by providing a

* Work conducted during an internship at AWS AI Labs.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4554–4570

survey of the current state-of-the-art for estimating the factual confidence in LLMs, and performing a systematic empirical comparison of the reliability and robustness of the existing methods.

We introduce an experimental framework, shown in Figure 1, enabling a fair comparison between methods across models and datasets. We adopt two distinct formulations for measuring factual confidence: (*i*) the probability of a statement to be true, noted $P(\text{True})$—fact-verification (Thorne et al., 2018; Azaria and Mitchell, 2023), and (*ii*) the probability of yielding a truthful answer to a query, noted $P(\text{I know})$—question answering (Kadavath et al., 2022; Yin et al., 2023). Additionally, we categorize the existing methods into five groups: trained probes, sequence probability, verbalization, surrogate token probability, and consistency-based.

Our experiments across eight publicly available LLMs indicate that prompting-based methods are less reliable than supervised-probing, although the latter requires training data and access to models' weights. For instruction-tuned LLMs, *verbalization* and *consistency-based* methods are viable alternatives. Further, we argue that all methods for estimating factual confidence can ultimately lead to misleading conclusions if only tested on a single way of asserting a fact: An LLM may seem to know a fact given an input, but then contradict itself given an alternative phrasing of that fact (Elazar et al., 2021; Kassner et al., 2021; Lin et al., 2022b; Qi et al., 2023; Kuhn et al., 2023). In our experiments, we find evidence of such instability, suggesting that LLMs do not always encode facts based on abstractions over diverse input variations.

In summary, this paper provides the following contributions:

- A survey of the literature on LLM factual confidence estimation;
- An experimental framework enabling a fair comparison across methods proposed in the literature;
- Insights about the reliability and robustness of different types of methods, providing recommendations for NLP practitioners;
- Insights about the consistency of factual confidence across semantically equivalent inputs.

2 Factual Confidence: Key Concepts

2.1 Definition of a Fact

We consider a *fact* to be a piece of information that accurately represents a world state.[1] A natural-language statement is *truthful*—or *factual*—if its meaning reports a state of affairs that is supported by a true fact: e.g., "*Paris is a city in France*" is truthful as the city of Paris is indeed located in France. Facts and natural-language statements are not linked by a one-to-one relation: The same fact can be declared with multiple statements, varying on the surface level, but sharing the same meaning.

For this reason, one's confidence in a fact should be consistent across meaning-preserving linguistic variations, such as paraphrases or translations of a statement: If we are certain that "*Paris is a city in France*" is true, we will not doubt that its paraphrase "*Paris is a French city*" or its translation in French (if we understand French) are also true.

2.2 Factual Confidence

We distinguish between two facets of factual confidence of LLMs, following Kadavath et al. (2022):

$P(\textbf{True})$ – shortened as $P(\text{T})$: the degree to which a model considers likely that a fact stated in the input is true. For example, "*Paris is the capital of France*" should get a high $P(\text{T})$ as it is truthful and is common knowledge, while "*Sidney is the capital of France*" should get a low $P(\text{T})$. To estimate $P(\text{T})$ scores we need to pass a statement in the input, which is evaluated in its truthfulness: this is in line with the setup of **fact-verification** (Thorne et al., 2018; Hardalov et al., 2022; Guo et al., 2022).

$P(\textbf{I Know})$ – shortened as $P(\text{IK})$: the degree to which a model considers likely that it will return the correct answer to an input querying about a fact. For instance, we can compute $P(\text{IK})$ in a **QA** setup passing a question as input—e.g., "*What is the capital of France*?". If confident to know the true answer, $P(\text{IK})$ should be high; it should instead be low in case of uncertainty. In contrast to $P(\text{T})$, $P(\text{IK})$ is estimated without stating the fact in the input, but rather expecting a factual answer by the model complementing the query.

$P(\text{T})$ and $P(\text{IK})$ are both telling of the underlying factual confidence of an LLM. However, depending on the data format (e.g., statements vs. questions) or task of interest (e.g., fact-verification vs. QA) focusing on one of the measures is more suitable. Previous works introducing methods to estimate factual confidence have

[1] For simplicity, in this work, we restrict our focus to minimal, atomic facts, in the sense that they do not involve a combination of other facts; e.g., "*The Louvre is in Paris*" as opposed to "*The Louvre is in Paris, which is in France*".

typically addressed only one of the two measures. However, as we demonstrate with our experimental framework, most methods can be adapted to estimate both $P(\text{T})$ and $P(\text{IK})$, although in practice they may not be equally reliable in each setup.

2.3 Robustness of Factual Knowledge

We work from the hypothesis previously voiced by Petroni et al. (2019) that a language model's factual knowledge may stem from encoding facts in its weights (*parametric memory*) as an abstraction over the linguistic input in the training data.

Such human-like robustness and abstraction ability cannot however be taken for granted (Bender and Koller, 2020; Mitchell and Krakauer, 2023; Mahowald et al., 2024). Testing for consistency to meaning-preserving variations of an input is key to distinguishing whether a model has encoded a fact as an abstraction over linguistic forms, as opposed to memorizing statements asserting the fact (Carlini et al., 2023). For instance, if a model has a robust encoding in its parametric memory of what the capital of France is, it should provide the same answer to "*What is the capital of France*?", "*What is the name of the French capital city*?" or any other rewording. Prior works already provided evidence that models may not always act consistently across semantically equivalent inputs (Elazar et al., 2021; Kassner et al., 2021; Ohmer et al., 2023; Qi et al., 2023). However, this has not yet been investigated in relation to the degree of factual confidence.

3 Factual Confidence: Survey of Methods

Based on a review of the research area, we identify five groups of existing methods to estimate factual confidence, which we discuss in the following subsections. In Table 1, we provide an overview of the functional differences among these methods.

3.1 Trained Probes

The methods in this group are based on probes which compute a transformation of a model's internal representations. The probes can take as input the final layer or earlier layers, taking advantage of the latent compression stages of an LLM (Voita et al., 2019).

Azaria and Mitchell (2023) proposed to train multi-layered probes to extract factual confidence scores from hidden states, under the argument that such estimates are less subject to surface-level features—how a claim is phrased—and thus more reliable. Their setup is in line with an estimate of $P(\text{T})$. Kadavath et al. (2022) adopted this method in a QA setup, estimating $P(\text{IK})$ using a trained value head on top of the final transformer layer. Breaking off from the constraints of supervised training, Burns et al. (2023) propose another version of the probe, which they train in an unsupervised manner, by maximizing distance between representations of contradicting answers on a Yes/No question dataset. They report performance only slightly lower than supervised alternatives—we therefore do not separate it in its own group. This method is worth considering in the specific case where a model's layer outputs are available but reliable annotations are not available to train a probe.

We must emphasize that these methods have more strict requirements compared to the other groups, as: (*i*) they require having access to the model weights, and (*ii*) they need supervised training data, i.e., data with labels that reflect if the statements are true, for $P(\text{T})$, and data with labels that reflect whether the model will provide the correct answer to the question, for $P(\text{IK})$.

3.2 Sequence Probability

This family of methods use the averaged log probabilities, assigned to a sequence of output tokens, to estimate factual confidence. For this approach to work well, sequence probabilities need to be well calibrated (Guo et al., 2017) to indicate probability of correctness (Fomicheva et al., 2020; Xiong et al., 2024). In the context of factual knowledge, sequence probability has been applied both in cloze tasks and QA setups (Jiang et al., 2021; Yin et al., 2023), which corresponds to measuring $P(\text{IK})$.

A major limitation of the sequence probabilities is that they represent the confidence over *how* a claim is made (i.e., the probability of the generated response as a sequence of tokens—different surface realizations of the claim would have different probabilities), rather than the confidence about the claim itself (Lin et al., 2022a). Moreover, Gal and Ghahramani (2016) showed that the output sequence probabilities produce unreliable, overconfident estimates in general. Thereby, these methods are mainly used as a weak baseline for model confidence estimation.

3.3 Verbalization

In the verbalization (*verbalized confidence*) methods (Lin et al., 2022a; Xiong et al., 2024; Yin et al., 2023; Tian et al., 2023; Kadavath et al., 2022), the

	Black-box	Trained	Prompt-based	Scores for
Trained Probe	No	Yes	No	$P(\text{T})$ & $P(\text{IK})$
Sequence Probability	Yes (*)	No	No	$P(\text{T})$ & $P(\text{IK})$
Verbalization	Yes	No	Yes	$P(\text{T})$ & $P(\text{IK})$
Surrogate Token Probability	Yes (*)	No	Yes	$P(\text{T})$ & $P(\text{IK})$
Consistency	Yes	No	No	$P(\text{IK})$

Table 1: Differences across types of factual confidence estimators. *Black-box* marks methods which do not rely on access to model's weights; (*) denotes the possibility to use sampling if token probabilities are not available.

model is directly prompted to report its confidence level (e.g., "*How confident are you that the answer is correct*?"). Here, differently from the other methods, the LLM generates a numeric confidence level as a sequence of output tokens. Lin et al. (2022a) and Tian et al. (2023) find that this method provides well-calibrated and surprisingly accurate estimates for highly capable instruction-tuned models like GPT-3 (Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023). Tian et al. (2023) extend the verbalization to "top-k" prompting, i.e., prompting the model for k answers, along with their estimated confidence. They also propose to use Chain-of-Thought (CoT) and multi-step prompting, e.g., first providing an answer and then providing a measure of confidence that this answer is correct. In this work, we focus on the simplified non-CoT, $k = 1$ prompting as it shows competitive results to the more complex methods for many model-dataset-metric combinations.

3.4 Surrogate Token Probability

These methods, extensively studied by Kadavath et al. (2022); Xiong et al. (2024), can be considered a hybrid approach between the *Sequence Probability* (Section 3.3) and *Verbalization* (Section 3.2). The input prompt asks the model to provide as output specific tokens to report the factuality of the claim in the input; the probabilities assigned to those tokens are then used to determine the confidence level. This method can be adapted to measure both $P(\text{T})$ and $P(\text{IK})$ (Kadavath et al., 2022).

3.5 Output Consistency

Output consistency methods (also referred to as *self-consistency*) (Wang et al., 2023b) build on the assumption that the high confidence of a LLM leads to generating consistent outputs. Given a question or an incomplete statement, we sample multiple completions and take the inter-responses consistency as the confidence measure. If the model consistently generates the same answer, the confidence is high. Conversely, if the model generates contradictory answers, the confidence is low. One limitation of this method is that, due to its dependence on completion generation, it can only be used to estimate $P(\text{IK})$ and not $P(\text{T})$.

Manakul et al. (2023) demonstrated the efficacy of this method when applied to factual knowledge, focusing on GPT models and using output consistency to validate model responses. Kuhn et al. (2023) argued that additional clustering of the outputs that are semantically equivalent as instances of the same answer is needed.

4 Methodology

4.1 Data

Currently, there is no standardized set of benchmarking datasets that are adopted across previous work. To fill this gap, we adapt two publicly available datasets, namely Lama T-REx (Petroni et al., 2019) and PopQA (Mallen et al., 2023), to test factual confidence in both fact-verification ($P(\text{T})$) and QA ($P(\text{IK})$) setups. We believe that establishing such a baseline setup is important both for researchers and practitioners.

4.1.1 $P(\text{T})$ in Fact Verification: Lama T-REx

Lama T-REx (Petroni et al., 2019) is a relational dataset made of triplets extracted from Wikipedia <subject, relation, object>, (e.g., <*Victor Hugo, was born in, France*>). We use this dataset to create both true and false statements for estimating $P(\text{T})$. We create false versions of each factual statement, by randomly substituting the object in the triplet with one from the same relation ("Victor Hugo was born in China"). This ensures the right entity type and avoids grammatical errors.

There are 34K triplets (true statements) in the T-REx dataset. For each true statement we sample one false, creating a balanced set of 50/50 true/false

statements. Then, we take 80% of the examples, 27K of the positives and their corresponding negatives, 54K in total, for training (only used for *trained probe*). The rest, we use for analysis—6.8K T-REx true statements and an equal number of false statements, 13.6K in total.

4.1.2 $P(\text{IK})$ in QA: PopQA

The PopQA dataset (Mallen et al., 2023) consists of short questions and single entity answers (e.g., question: *What is George Rankin's occupation?*, answer: *Politician*.). This dataset offers a set of synonymous phrases for each correct answer. This is an important feature, which makes the evaluation more robust to answer phrasing, and, in turn, lowers the risk of underestimating model's correctness. Moreover, this dataset covers a broad range of entities, with varying degrees of popularity (estimated based on the number of Wikipedia page views). On one hand, this ensures the diversity of the target entities, and, on the other, it allows for further analysis between popularity and estimated confidence, which we leave for future work.

We use PopQA to test models' factual confidence given a fact-related query, i.e., $P(\text{IK})$. The dataset contains 14K questions: we keep 80% (11K) for training, and 20% (2.8K) for testing. By definition (Section 2.2), the gold labels for $P(\text{IK})$ should indicate if the model outputs a correct answer. Ultimately, a model's answer depends on the decoding strategy; in this work, for simplicity and clarity of interpretation, we use greedy decoding. If the answer is correct, we set the gold $P(\text{IK})$ to 1, else to 0 (more details in Section 4.2). As the labels depend on model correctness, the data will have varying proportions of positive labels across models, in our case, ranging from $\sim$11% to $\sim$27%.[2]

4.2 Scoring Methods Implementation

Below, we report the main specifics of our implementation of the methods (details in Appendix A).

4.2.1 Estimating $P(\text{T})$

Given a statement, we compute $P(\text{T})$ as follows:

Trained probe: As in Azaria and Mitchell (2023), we train a 3-layer fully connected feed-forward network for 10 epochs, passing as input the hidden states from the 24th transformer layer for each LLM (requires hyper-parameter optimization, later layers before the last one work better) and predicting whether a statement is true or false.

Sequence Probability: Average log-probability of the statement's tokens.

Verbalization: Prompting for the confidence level that the statement is true (Appendix A).

Surrogate Token Probability: Log-probability of the "Yes" token following a query on whether the statement is true.

4.3 Estimating $P(\text{IK})$

Below, we describe how the $P(\text{IK})$ estimates for each method group are computed, based exclusively on the question.

Trained Probe: We use the same approach as for $P(\text{T})$, but train the probes to predict whether the model's greedy-generated answers will be truthful or not.[3]

Sequence Probability: Average log-probability of the question's tokens.[4]

Verbalization: Prompting for the confidence level of knowing the answer to the question (see Appendix A for details).

Surrogate Token Probability: Log-probability of "Yes" token following a query on knowing the answer to the question.

Consistency: We prompt the model with the question and sample 10 responses with $\tau = 1$. Then, we compute a matrix of pairwise NLI scores (Laurer et al., 2024) on all generations, and return an average.

4.4 Evaluating Scoring Methods

To evaluate the methods, we use the area under the precision-recall curve (AUPRC), as is common in related literature, e.g., Kadavath et al. 2022. Using a metric that considers various decision thresholds enables a robust comparison across methods. The higher AUPRC, the better ranking capability of the method, with cleaner separation between true/false statements or known/unknown facts. In an effort to make interpretation of AUPRC more intuitive for practical applications, we also report the precision and recall at K (Tables 5 and 6 in Appendix D).

[2] The questions from PopQA are generally considered hard (ChatGPT: 30% accuracy, SelfRAG (Asai et al., 2024): 55%).

[3] This is a simpler, less computationally expensive version of the approach of Kadavath et al. (2022), where multiple answers are sampled and the probe initially predicts a continuous score—proportion of correct answers in the sampled set.

[4] This implementation captures how surprised the model is by the question, which is linked with expected correctness.

Names	Size	Open	Arch.	Instruct
Falcon	40B	✔	Dense	
Falcon Inst.	40B	✔	Dense	✔
Falcon	7B	✔	Dense	
Falcon Inst.	7B	✔	Dense	✔
Mixtral	46.7B	✔	SMoE	
Mixtral Inst.	46.7B	✔	SMoE	✔
Mistral	7B	✔	Dense	
Mistral Inst.	7B	✔	Dense	✔

Table 2: The models used in our experiments. *Dense* represents the usual transformer decoder architecture, while SMoE stands for Sparse Mixture of Experts (Shazeer et al., 2017), here made of 8 experts of 7B. *Instruct.* models have been instruction fine-tuned. Open models have publicly available weights.

4.5 Models

We study eight publicly available LLMs, with open access to weights. We consider models with different sizes (7B to 46.7B), architecture, and training paradigms (instruction-fine-tuned or not) from the Falcon (Almazrouei et al., 2023) and Mistral (Jiang et al., 2023, 2024) model families (see Table 2).

4.6 Paraphrasing and Translation

To test methods robustness and to disentangle confidence over a fact from confidence based on a specific wording, we generate semantically equivalent variants of statements/questions from Lama T-REx and PopQA (see Section 6). For each input, we generate 10 paraphrases by prompting Mixtral-8x7B-Instruct-v0.1 (prompt and examples are in Appendix B). We remove repetitions and only keep paraphrases that are semantically equivalent[5] to the original input. This results in an average of eight paraphrases per original input.

We also consider translation as another meaning-preserving transformation. Specifically, we translate the English examples to two languages from different language families (Romance and Slavic):[6] (*i*) French—a high resource language to which all models except the 7B Mistral models have been explicitly exposed to during training, and (*ii*) Polish—a language on which we expect lower degree of competence (Falcon reports "limited capability" for Polish (Almazrouei et al., 2023), while Mistral models do not mention it at all). Finally, we manually verified the quality of a sample of 100 translations, finding them to be meaning preserving and vastly without errors.

[5]To detect semantic equivalence, we test for entailment in both directions using an NLI model (Laurer et al., 2024)

[6]For translations we use Amazon Translate.

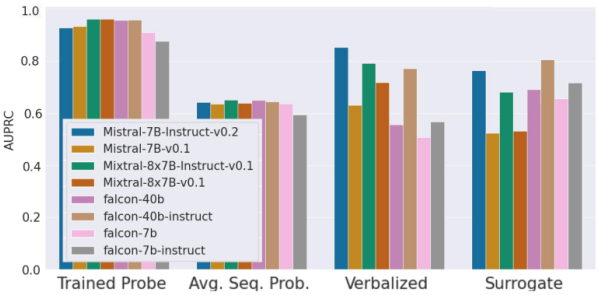

Figure 2: AUPRC scores on T-REx with both true and false statements; $P(\mathrm{T})$.

5 Empirical Comparison of the Methods

5.1 $P(\mathrm{IK})$ on Lama T-REx

With each of the four methods, we derive estimates of factual confidence for all statements in the Lama T-REx test set, repeating the experiment for each LLM. We evaluate the reliability of a method by checking whether it yields $P(\mathrm{T})$ scores that can effectively separate the true statements from the false statements, measured as AUPRC.

We report the results of this analysis in Figure 2. The *trained probe* method performs best, outperforming the *sequence probability* by an average AUPRC of .3. Of all methods and models, only the *verbalized method* is competitive to the supervised probe, and only for Mistral 7B instruct. Otherwise all methods perform at least .1 AUPRC below the *trained probe*. This result suggests that information about the expected truth value of a statement is better captured in deeper layers of the network, as opposed to the output scores.

While for *trained probe* and *average sequence probability* we note relatively small differences in AUPRC across models, for the *verbalized* and *surrogate* methods we see large variation. Concretely, instruction-tuned models always perform better than their counterparts. This is expected as both methods require to follow instructions in the prompt. Model size also seems to have an effect: all 40B+ models perform better than their 7B counter-parts, with the exception of Mistral-7B-Instruct-v0.2. Information on the specific differences between versions of Mistral is not publicly available, making the interpretation of this result difficult. One possibility is that the second version of the model has better instruction-following capa-

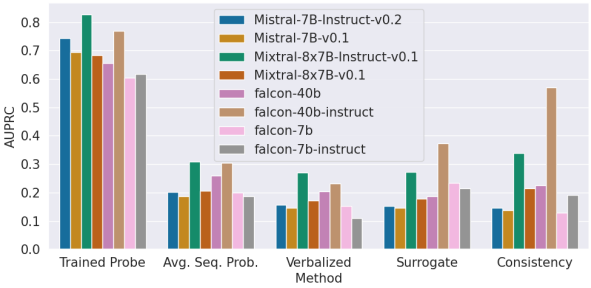

Figure 3: AUPRC scores on PopQA dataset; $P(\text{IK})$.

Name	Size	AUPRC	Δ
Falcon	40B	.80	-.16
Falcon Ins.	40B	.81	-.15
Falcon	7B	.66	-.25
Falcon Ins	7B	.59	-.28
Mixtral	46.7B	.78	-.18
Mistral	7B	.62	-.31
Mistral Ins	7B	.75	-.18

Table 3: AUPRC on PopQA test set re-worked as true/false statements, using $P(\text{T})$ estimates from probes trained on Lama T-REx. Δ: difference of AUPRC with respect to that for Lama T-REx data (in-domain).

bilities, but we cannot be sure (BehnamGhader et al. 2024 discusses conjectures on unusual properties of the Mistral family of models).

Finally, while the *average sequence probability* method performs consistently above chance (50%), it has overall poor performance in comparison to other methods. It only outperforms the other non-trained methods—*verbalized* and *surrogate*—for non-instruction-tuned models.

5.2 $P(\text{IK})$ on PopQA

$P(\text{IK})$ estimates the degree of a model's confidence that its predicted answer will be correct. A good estimator of $P(\text{IK})$ would thus assign high scores to queries which the model answers correctly, and low scores to others. Following this reasoning, for $P(\text{IK})$ we compute the AUPRC scores using binary labels that encode whether the model's answer (in our case, generated with greedy decoding) is correct. Note that this way of computing AUPRC—based on *future correctness*—provides a direct estimate of the method's expected effectiveness for hallucination mitigation (when the method is used to automatically detect when the model should abstain from answering). In this scenario, a method is effective only if its estimates are actually predictive of the correctness of the model's answers.

The results are reported in Figure 3. In this experiment we also study the *Consistency* method, which was omitted from $P(\text{T})$ results because, by design, it cannot be applied to an entire statement. Overall, $P(\text{IK})$ is harder to estimate than $P(\text{T})$, with lower AUPRC results: e.g., The best trained probe is 0.1 below in AUPRC for $P(\text{IK})$ than it is for $P(\text{T})$. This may be due to the complexity of the setup—in QA the confidence is estimated only based on a query, in contrast to fact-verification. But it may also be that the binary *future correctness* labels used for our AUPRC computation introduce some noise. E.g., the model may be genuinely uncertain and still output the correct answer by chance.

The *trained probe* method is again, by large, the most reliable across all models. With the exception of Falcon-40B instruct, the other methods perform close to or below chance (depending on the model's label distribution, chance level varies between 0.11 and 0.27). This indicates that non-trained estimators are generally not reliable for $P(\text{IK})$ despite being frequently used in the literature. Within each method, we observe differences across models—up to a 40% margin. This can be linked to (*i*) whether a model is instruction-tuned (as noted for $P(\text{T})$) and (*ii*) the model family—with more reliable scores for Mistral models than for Falcon models.

5.3 Generalization of the Trained Probe

The results above highlight the trained probe as the most reliable estimator for factual confidence—both for $P(\text{T})$ and $P(\text{IK})$. However, in those experiments we trained and evaluated the models within the same domain, which leaves open questions about the probe's generalization capabilities. We address this gap by evaluating the model from 5.1, trained to estimate $P(\text{T})$ from Lama T-REx data, on the PopQA dataset converted to test for $P(\text{T})$. Specifically, we re-work the PopQA data for the fact-verification setup by turning question-answer pairs into (evenly distributed) true and false statements, using the template: "*The answer to [QUESTION] is [ANSWER]*".[7] We derive estimates for $P(\text{T})$ on such statements using the probes trained on Lama T-REx, and compute AUPRC (Table 3).

Going from in-domain to out-of-domain test data

[7]For true statements we use the gold answers from PopQA dataset. For false statements, we sample alternative answers from the same question class in the dataset; e.g., *The answer to* "In which country is Washington?" *is* "United States of America" vs. "*South Korea*".

(Lama vs. PopQA), we observe AUPRC differences of min -.15 and max -.31. However, the scores remain in a high range of [.62, .81] indicating substantial generalization. The LLMs for which the probe retains the least and the most reliability are Mistral-7B and Falcon-40B-instruct, respectively. Interestingly, these are also the models getting the least and the most answers right on PopQA in the QA setup (14% and 23%). This suggests that the transferability of the probe may be affected by how challenging the out-of-domain dataset is to the model. In the next sections, we provide further evidence of probe generalization by looking at whether and to what extent the AUPRC is affected by input paraphrasing and translation.

6 Robustness to Linguistic Variations

In this section, we apply meaning-preserving linguistic variations to each input in order to: (*i*) Assess the robustness of methods. The expectation is that if a method is robust, it should produce equally reliable estimates (equally high AUPRC scores) across different input formulations (Section 6.1); (*ii*) Investigate the stability of an LLM's encoding of facts. The expectation is that if a fact is well abstracted, the factual confidence should be invariant to semantics-preserving changes in the input (Section 6.2). We consider two types of input variation: paraphrases and translations.

6.1 Robustness of Methods

We study method robustness in both $P(\text{T})$ and $P(\text{IK})$, using the same setup as before; in particular, we do not retrain the *trained probe* and do not adapt the prompts in any way.[8] To test robustness on paraphrases, we generate 10 different paraphrase sets—each holding different formulations of the original inputs—and compute AUPRC on each set. We observe that AUPRC remains stable for all methods (absolute variation between 5% and 10%), indicating they are robust to paraphrasing. The most affected method is the *trained probe* in the $P(\text{IK})$ setting, but even here we only note up to a standard deviation of 3 percentage points (for Mistral-7B). Full result tables are in Appendix C.

For translations, we compute a separate AUPRC on the French and Polish versions of T-REx. We find varying degrees of method transferability to new languages (above .5 for *verbalized*, up to .91 for *trained probe*). All methods generalize to both French and Polish above chance, except for (*i*) *verbalized confidence* and (*ii*) *surrogate logits* when applied to Mistral models (see Figure 7 in the Appendix). Notably, the probes trained on English data remain to a large extent reliable (AUPRC for French: .73-.91; for Polish: .61-.91) on unseen languages—with 40B+ models and the instruction-tuned Mistral demonstrating the highest transferability. This provides additional evidence for out-of-domain generalization of trained probes (Section 5.3). In particular, the probes can extract scores that are discriminative of true and false statements also from hidden states computed from inputs in a different language than the one used at training. This suggests that the LLMs encode factual confidence in a similar way across languages, which is remarkable, given the differences in exposure to training data from the three languages.

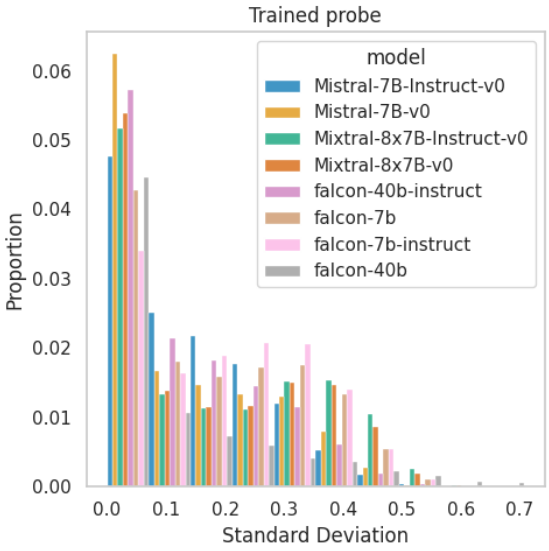

Figure 4: Distribution of standard deviation scores for normalized $P(\text{T})$ on paraphrases of the same fact.

6.2 Robustness of Facts Encoding in LLMs

We hypothesize that, to robustly learn facts and minimize hallucinations, a model has to build stable abstractions over different types of relevant evidence from the training data. We also expect that if the model has built such a robust representation of a fact, this would lead to equal confidence under equivalent formulations of that fact. Inconsistent confidence would in turn indicate excessive reliance on surface-level features.

Figure 4 shows how paraphrasing the input (∼8 paraphrases per input causes changes in the *trained*

[8] Note this also applies to translations; i.e., the trained probe is trained on English data only and we use English prompts to query the model about French/Polish inputs.

probe $P(\text{T})$ estimates across the T-REx dataset.[9] Specifically, we present the distribution of the standard deviation of confidence scores, across different facts. The amount of variation is not stable across facts. While for each model, we observe a large amount of facts for which there is little-to-no variation in the score (indicating a stable encoding), we also note many facts for which different wordings lead to strongly varying degrees of confidence (up to .5 standard deviation, with very few cases at .7). This indicates inconsistent LLM behavior, with excessive sensitivity to how a claim is worded. Out of all models, Falcon-7B-instruct appears to have developed the least stable encoding, with the standard deviation distribution shifted towards higher values. On the other extreme, Mistral-7b-v0 appears to be the model showing less variation.

To test robustness of factual knowledge across languages, we compare the distributions of $P(\text{T})$ scores over the same facts using the Spearman correlation analysis (for language pairs) and the Friedman test (for language groups). Analysis reveals high correlations (Spearman's $\rho > .7$; full results in Table 4 in Appendix) between factual confidence scores on all language pairs for the 40B+ models. In particular, we note the highest correlations (in the .87-.92 range) for Falcon 40B models, which points to highly robust multilingual behavior. However, the Friedman tests reveal that for all models, the differences across the distributions are statistically significant (p-values close to 0); i.e., the differences in scores across the languages are not close enough to be coming from the same population. Given those results, we conclude that while there is a link between the confidence scores across the languages, this is not fully systematic.

7 Discussion & Conclusion

In this paper, we compare existing methods to estimate LLMs factual confidence. Obtaining reliable estimates can benefit LLMs applications, by anticipating potential hallucinations and limiting the non-factual information output by a model (Tonmoy et al., 2024; Evans et al., 2021). However, if not reliable, such estimates can be counterproductive, as they would introduce errors and negatively affect user-model interactions.

Our experiments across eight LLMs demonstrate that the *trained probe* method is the most reliable estimator of LLM factual confidence. It works well for both fact-verification ($P(\text{T})$) and Question Answering ($P(\text{IK})$) across all 8 tested models, indicating that its reliability is likely to generalize to other LLMs. Moreover, we show that it generalizes to out-of-domain data: (*i*) when a model trained on T-REx is applied to an unseen dataset (PopQA, Section 5.3), and (*ii*) in a cross-lingual transfer setting (Section 6.1). We must note, though, that the fine-tuning nature of the method clearly puts it at an advantage over the other zero-shot probing methods. Moreover, applying the *trained probe* method comes with strong requirements: (*i*) access to model weights—not always provided by proprietary LLMs, and (*ii*) need of supervised data. If these requirements cannot be met, but the model is instruction-tuned (Ouyang et al., 2022) we recommend estimating $P(\text{T})$ with *verbalized confidence* or *surrogate probabilities*. The other methods under study, especially if applied to non-instruction-tuned LLMs, are not consistently reliable.

Our results highlight the need for more research to develop reliable estimators that can be applied to black-box models, with inaccessible internal representations. We expect the reliability gap between methods like *verbalized confidence* and *trained probe* to get smaller with increasingly powerful LLMs, especially in their ability to follow instructions. However, strong results of *trained probe* indicate that hidden states contain signal about factual confidence, and it is unclear whether this can ever be fully leveraged by the prompting approaches.

Besides the comparison among methods, we also provide insights on the stability of factual knowledge in LLMs (Petroni et al., 2019; Mitchell and Krakauer, 2023; Mahowald et al., 2024). We show that the factual confidence of an LLM is not always consistent under meaning-preserving variations of the input (paraphrases and translations). While the model may sometimes be sure that a fact is true or false, or that it knows the answer to a question, it may actually behave differently if we reformulate the statement or question. An interesting direction for future research is the exploration of training methods that teach an LLM to better disentangle facts from the diversity of forms they can be stated in, and ultimately exhibit better and more consistent factual knowledge. This would also contribute to increasing LLMs resistance to adversarial attacks (Madry et al., 2018), mitigating the generation of misinformation due to an incorrect sensitivity to input changes.

[9] We focus on the *trained probe* since it was the best performing method.

Acknowledgments

We thank the anonymous reviewers for their helpful questions and comments, which have helped us improve the quality of the paper. We also want to thank Ionut Sorodoc, Neha Anna John, Phu Mon Htut and the entirety of the AWS Bedrock Responsible AI team for the time they took to support, discuss and improve this work. We thank Marco Baroni and Nathanaël Rakotonirina for helpful discussion during this project.

Matéo Mahaut is supported by the European Research Council (ERC) as part of a project in the European Union's Horizon 2020 research and innovation program (grant agreement No. 101019291). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Limitations

Given the extensive scope of this work (eight models, five methods and two facets of factual knowledge), we did not have the capacity to study more complex aspects of factual confidence, such as non-atomic facts, reasoning or in-context learning. While our results show that the *trained probe* is much stronger than other methods on T-REx and PopQA, there is no guarantee that this remains the case in more complex settings. Furthermore, methods themselves have limitations, making comparison use-case dependent. The *trained probe* method for example requires training data, and while we have tested for transfer capabilities in our simple atomic fact setup, Kadavath et al. (2022) have shown that there are limits to the kind of tasks this method can be transferred to. The same can be said of the *sequence probability* method, which in our experiments works better than both prompt-based methods for non instruction fine-tuned models. While this method performs well on simple atomic facts, more complex sentences, or even simple but longer sentences could lead to weaker results. Furthermore, both prompt-based methods are sensitive to prompt-variations.

Ethics and Broader Impact

This work contributes to the wider goal of automatically reducing risk when using LLMs. We contribute to false statement detection and answer confidence, leading to potential applications which can build trust in LLMs. None of the methods studied completely solve the issue of hallucination, or non-factual utterances of models, leaving a need for future works on the subject. While methods studied can work with models with 7B and 40B+ parameters, the deployment of those models requires specific infrastructure and is compute intensive.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, ICLR '24, Vienna, Austria.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings '23, pages 967–976, Singapore.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 5185–5198, Online.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, ICLR '23, Kigali, Rwanda.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, ICLR '23, Kigali, Rwanda.

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.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *ICML '16*, pages 1050–1059, New York City, NY, USA.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML '17'*, pages 1321–1330, Sydney, Australia.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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*, EACL '21, pages 3250–3258, Online.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, ICLR '23, Kigali, Rwanda.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pages 3214–3252, Dublin, Ireland.

Linhao Luo, Trang Vu, Dinh Phung, and Reza Haf. 2023. Systematic assessment of factual knowledge in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings '23, pages 13272–13286, Singapore.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, ICLR '18, Vancouver, Canada.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, pages 1364–6613.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ACL '23, pages 9802–9822, Toronto, Canada.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 9004–9017, Singapore.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics*, GEM '23, pages 258–276, Singapore.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS '22*, pages 27730–27744, New Orleans, Louisiana, USA and Online.

Lorenzo Pacchiardi, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Yue Pan, Yarin Gal, Owain Evans, and Jan M. Brauner. 2024. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In *The Twelfth International Conference on Learning Representations*, ICLR '24, Vienna, Austria.

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 '19, pages 2463–2473, Hong Kong, China.

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*, EMNLP '23, pages 10650–10666, Singapore.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations*, ICLR '17.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*, abs/2102.02503.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification*, FEVER '18, pages 1–9, Brussels, Belgium.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 5433–5442, Singapore.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the Transformer: A study with machine translation and language modeling objectives. 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 '19, pages 4396–4406, Hong Kong, China.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, ICLR '23, Kigali, Rwanda.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, ICLR '24, Vienna, Austria.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, Findings '23, pages 8653–8665, Toronto, Canada.

Appendix for "Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators"

A Implementation of scoring methods

A.1 Verbalized

For the *verbalization* method, we start with the prompt from Tian et al. (2023) and adapt it to the current tasks. Then, we filter out non-numerical outputs and normalize.
Prompt used for $P(\text{T})$ experiments:

> *Provide your confidence level (on a scale of 1.0 to 10.0) that the following statement is correct .*
>
> *The statement is: $statement*
>
> *Confidence level:*

Prompt used for $P(\text{IK})$ experiments:

> *Provide your confidence level (on a scale of 1.0 to 10.0) that you can answer the following question .*
>
> *The question is: $question*
>
> *Confidence level:*

A.2 Surrogate token probabilities

For $P(\text{T})$ we use the prompt:

> *Q: Is the following statement true: $statement (Yes/No/Maybe)?*
> *A:*

For $P(\text{IK})$ we use the prompt:

> *Q: Do you know the answer to the following question: $question (Yes/No/Maybe)?*
> *A:*

A.2.1 Consistency

The ground truth answer is generally a word, ranging between two and five tokens. We output 10 samples of up to 25 tokens using a temperature (τ) of 1, leaving ample room for the ground truth to be generated in a sentence. Finally, we calculate the consistency score by computing the mean NLI (Laurer et al., 2024) score of these 10 samples.

Name	Size	En-Fr	En-Po
Falcon	40B	.90	.86
Falcon Ins.	40B	.92	.87
Falcon	7B	.79	.44
Falcon Ins	7B	.67	.35
Mistral	7B	.67	.58
Mistral Ins	7B	.65	.53
Mixtral	46.7B	.87	.77

Table 4: Spearman correlation coefficient for English-French and English-Polish $P(\text{T})$ scores on translated Lama T-REx statements.

B Paraphrasing

Prompt used to generate paraphrases with Mixtral-8x7B-Instruct-v0.1 (examples are provided in Table 7):

> *Given a sentence, generate paraphrases of it as follows:*
>
> *- You can change and/or add words, and/or change the syntactic structure of the sentence;*
>
> *- Make sure the new sentence does not add additional details with respect to the original.*
>
> *- Make sure the new sentence does not omit any details with respect to the original.*
>
> *- Make sure the new sentence is natural and plausible, in spite of the changes.*
>
> *- Do not generate the original sentence or previously generated ones.*
>
> *List your paraphrases as bulletpoint.*
> *Sentence: $sentence*
> *New sentences:*

The variation of $P(\text{T})$ and $P(\text{IK})$ with paraphrases are provided in Figures 5 and 6.

C Method Robustness to Variation

To measure the robustness of the methods towards linguistic variations, we randomly sample a paraphrase for every sentence in the original dataset, making ten sets of paraphrases of the same size.

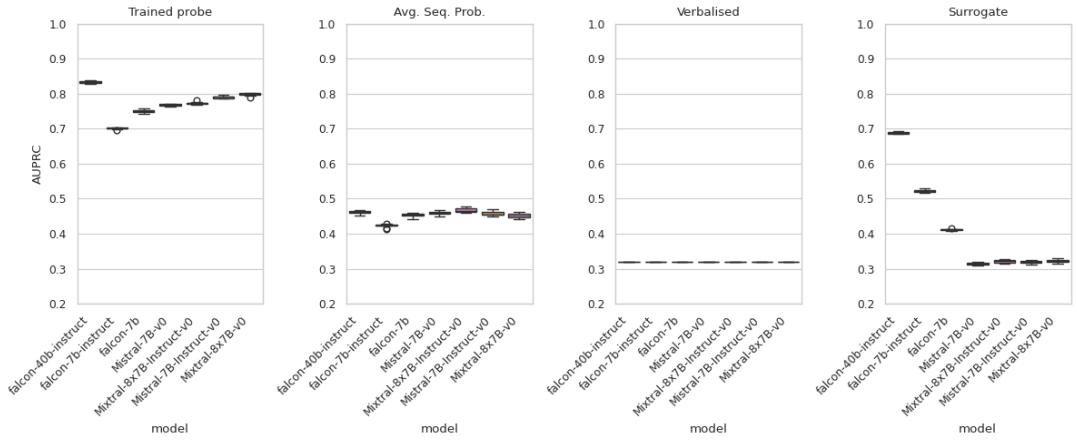

Figure 5: Variation in $P(\mathrm{T})$ AUPRC when sampling paraphrases. 10 sets of paraphrases are randomly sampled, with one paraphrase for every question in Lama Lama T-RE.

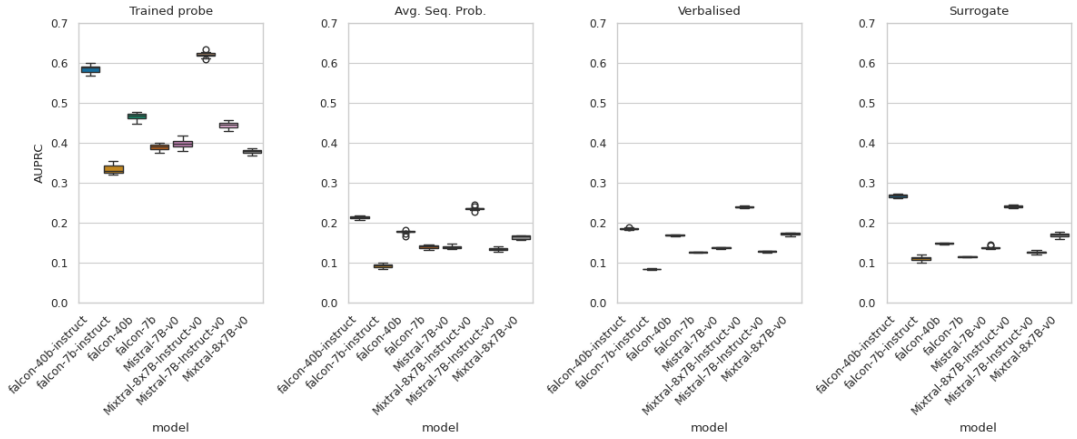

Figure 6: Variation in $P(\mathrm{IK})$ AUPRC when sampling paraphrases. 10 sets of paraphrases are randomly sampled, with one paraphrase for every question in PopQA.

We then compute AUPRC without changing the method in any way for the ten sets, and calculate the variance in results. The results are shown in Figures 5 and 6. All methods remain stable, and robust to paraphrases. The biggest variation occurs for the *trained probe method*, but are only of the order of 3 percentage points.

Table 4 shows the correlation between scores across different languages, and Figure 7 shows AUPRC of all four methods evaluated on the French and Polish versions of the Lama T-REx dataset.

D Analysis of specific Precision and Recall

In Tables 5 and 6 respectively, we show results for precision at different recall thresholds for the two formations: $P(\mathrm{T})$ on Lama T-REx and $P(\mathrm{IK})$ on PopQA. Similar to our previous findings (see Figures 2 and 3) the *trained probe* noticeably outperforms other methods, making the differences much more apparent, confirming that for $P(\mathrm{IK})$ the Falcon 40B outperforms all other models, and all methods, expected trained probe, show low performance when applied with Falcon 7b Instruct. In constrast to the other methods, the precision of the *verbalized confidence* does not change for different recall thresholds, for both $P(\mathrm{T})$ and $P(\mathrm{IK})$, which suggests that we are reaching its limits to estimate factual confidence. We hypothesize that the emphverbalized confidence failing to disentangle correctly *True* sentences from *False* ones. This is also true, to some extent, for the *surrogate token* method and the *sequence probability* for $P(\mathrm{IK})$. The *trained probe*, on the other hand, has better precision with lower recalls, which is also the case for the *sequence probability* for $P(\mathrm{IK})$, as well as the *surrogate token* method in the same condition,

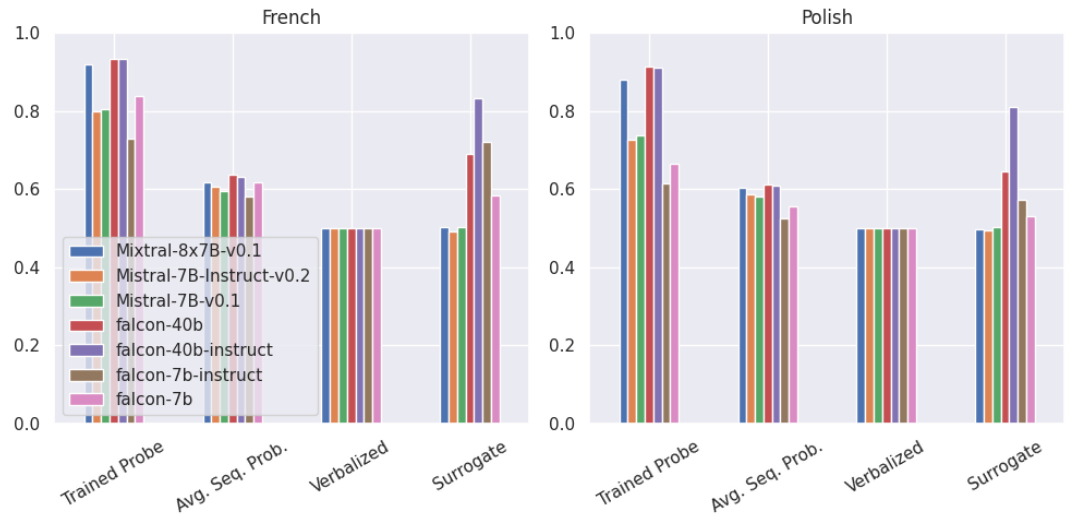

Figure 7: AUPRC for $P(\mathrm{T})$ scores for translations in French and Polish of the Lama T-REx statements.

Name	**Size**	**Surrogate**			**Trained Probe**			**Avg. Seq. Prob**			**Verbalized**		
		r90	*r70*	*r50*	*r90*	*r70*	*r50*	*r90*	*r70*	*r50*	*r90*	*r70*	*r50*
Falcon	40B	.33	.34	.45	.60	.77	.87	.36	.40	.45	.32	.32	.32
Falcon Ins.	40B	.43	.57	.69	.62	.78	.88	.36	.41	.46	.32	.32	.32
Falcon	7B	.32	.32	.33	.49	.64	.77	.36	.40	.44	.32	.32	.32
Falcon Ins.	7B	.34	.37	.41	.60	.76	.87	.34	.37	.41	.32	.32	.32
Mixtral	46.7B	.32	.32	.31	.59	.75	.84	.35	.39	.44	.32	.32	.32
Mixtral Ins.	46.7B	.32	.31	.30	.60	.74	.81	.36	.41	.46	.32	.32	.32
Mistral	7B	.32	.32	.32	.53	.67	.80	.35	.40	.44	.32	.32	.32
Mistral Ins.	7B	.32	.32	.32	.56	.71	.83	.36	.41	.45	.32	.31	.30

Table 5: Method precision for estimating $P(\mathrm{T})$ on the Lama T-REx dataset for 3 recall values (P@90, P@70, P@50). Here, we set a threshold to ensure a certain recall, and look at the resulting precision. A recall of 90 with .72 precision would mean that when we select a score threshold that ensures 90% of True sentences are correctly classified as such, 72% of all sentences in the tested dataset are correctly classified.

however only with bigger Falcon models. These results do push for additional work, as they point out that there remains substantial overlap when classifying True and False sentences (as well as successfully or unsuccessfully completed sentences) with a maximum of around 60% of sentences correctly classified, reaching a recall of 90 in both settings.

Name	Size	Surrogate			Trained Probe			Avg. Seq. Prob.			Verbalized		
		r90	*r70*	*r50*	*r90*	*r70*	*r50*	*r90*	*r70*	*r50*	*r90*	*r70*	*r50*
Falcon	40B	.33	.34	.45	.60	.77	.87	.36	.40	.45	.32	.32	.32
Falcon Ins.	40B	.26	.36	.40	.27	.37	.48	.17	.18	.18	.17	.17	.17
Falcon	7B	.11	.11	.11	.20	.26	.31	.11	.11	.12	.11	.11	.11
Falcon Ins.	7B	.07	.07	.07	.20	.26	.27	.07	.08	.08	.07	.07	.07
Mixtral	46.7B	.16	.16	.16	.18	.24	.29	.16	.15	.15	.16	.16	.16
Mixtral Ins.	46.7B	.21	.21	.22	.23	.28	.37	.21	.21	.21	.21	.21	.21
Mistral	7B	.12	.12	.12	.14	.21	.25	.12	.12	.12	.12	.12	.12
Mistral Ins.	7B	.11	.12	.12	.15	.19	.27	.11	.12	.12	.12	.12	.12

Table 6: Method precision for estimating $P(\text{IK})$ on the PopQA dataset at for 3 recall values (P@90, P@70, P@50)

Original sentence	Paraphrases
	Lama T-REx
Michie Mee is a actress by profession.	Acting is the profession of Michie Mee. Michie Mee makes a living as an actress. Michie Mee is a professional actress. Michie Mee is an actress in her profession. Michie Mee is an artist who acts for a living.
The Munsters was originally aired on Bravo network .	Bravo network was the first to air The Munsters. The Munsters was first shown on Bravo. The Munsters was first transmitted on Bravo. Bravo was the first network to air The Munsters. The Munsters was first broadcasted on Bravo.
	PopQA
What is George Rankin's occupation?	What does George Rankin do for a living? What line of work is George Rankin in? What is George Rankin's job? What is George Rankin's profession? Can you tell me what George Rankin does? George Rankin's employment, could you tell me about it? George Rankin's work, what is it?
In what city was Louis Renault born?	Where did Louis Renault come into the world? In which urban area did Louis Renault enter the world? In what metropolis did Louis Renault make his appearance? In which city did Louis Renault first see the light of day? In which city was Louis Renault given birth?

Table 7: Examples of automatic paraphrasing from the T-REx and PopQA datasets.