# *Confidently Wrong*: Exploring the Calibration and Expression of (Un)Certainty of Large Language Models in a Multilingual Setting

**Lea Krause Wondimagegnhue Tsegaye Tufa**
**Selene Báez Santamaría Angel Daza Urja Khurana Piek Vossen**
Vrije Universiteit Amsterdam
{l.krause, w.t.tufa, s.baezsantamaria, j.a.dazaarevalo, u.khurana,
p.t.j.m.vossen}@vu.nl

## Abstract

While the fluency and coherence of Large Language Models (LLMs) in text generation have seen significant improvements, their competency in generating appropriate expressions of uncertainty remains limited. Using a multilingual closed-book QA task and GPT-3.5, we explore how well LLMs are calibrated and express certainty across a diverse set of languages, including low-resource settings. Our results reveal strong performance in high-resource languages but a marked decline in performance in lower-resource languages. Across all, we observe an exaggerated expression of confidence in the model, which does not align with the correctness or likelihood of its responses. Our findings highlight the need for further research into accurate calibration of LLMs especially in a multilingual setting.

## 1 Introduction

Accurate estimation of one's own certainty and the confidence in provided information is pivotal not only for humans but also for machine learning models, particularly those intended for broad use. When someone asks a person a factual question, they search in their memory for the answer and produce it when they find it. If they are not able to find it, they reply *I don't know*. It is also the case that the more hesitant they feel about knowing a fact, the more they fill their answers with hedges such as *I guess* or *uhm* to express they are not certain (Smith and Clark, 1993). Ideally, a capable LLM should be able to express its uncertainty about the facts for which it has weaker evidence when generating an answer. A significant body of research has examined the phenomenon of overconfidence in humans, shedding light on their tendency to overestimate their knowledge and capabilities (Lichtenstein and Fischhoff, 1977; Brenner et al., 1996). In parallel, a substantial amount of work has been dedicated to the calibration of models in machine learning, typically with a focus on English-language data (Guo et al., 2017; Si et al., 2022; Chen et al., 2023).

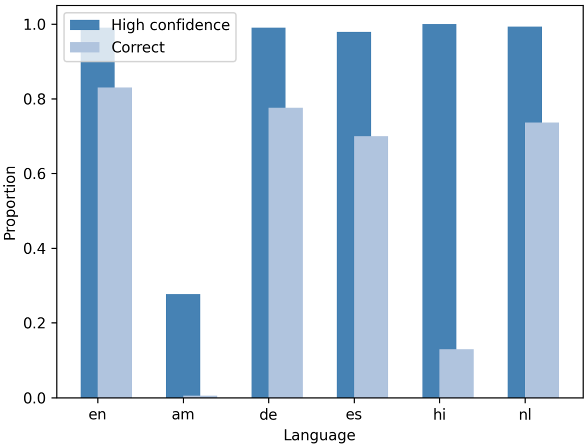


Figure 1: Proportion of *High Confidence* (HI) answers and *Correct* answers for each language. For Amharic, all non-HI answers (72.3%) were unintelligible (NI).

However, if multilingual support is stated when releasing an LLM, claims about model performance based solely on English datasets are not enough, and the capacity to show accurate performance across multiple languages is essential. As such, we investigate the *calibration*, *performance*, and *verbal expression of certainty* in an LLM (GPT-3.5) in a closed-book QA setting across six languages: Amharic, Dutch, English, German, Hindi, and Spanish, for which we release our dataset and annotations[1]. Our methodology for this investigation is based on a similar exploration conducted by Mielke et al. (2022), albeit their research was solely based on English-language performance.

Our findings reveal that while the performance of the model on high-resource languages is commendable, a significant drop in performance is observed in lower-resource languages. Interestingly, our results further demonstrate an excessively confident model, with confidence levels appearing to be unrelated to the correctness and, in most cases,

[1] https://github.com/lkra/multilingual_confidence_calibration

*Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 1–9
September 12, 2023

to the likelihood of its responses. Highlighting the need for continued research and improvement of calibration of language models, particularly in a multilingual setting.

## 2 Related Work

**Calibration and expression of confidence** In NLP, sequence generation poses a special challenge for understanding and estimating uncertainty. Malinin and Gales (2022) developed an ensemble-based probabilistic framework, while Kuhn et al. (2023) proposed the concept of semantic entropy as a measure of uncertainty in LLMs. Additionally, Si et al. (2022) introduced MacroCE, a novel calibration metric for LLMs.

Efforts to improve LLM reliability have been introduced by studies such as Si et al. (2023), which devised prompting strategies for GPT-3, and Kadavath et al. (2022), who explored LLM self-evaluation. Finally, the interpretation and generation of uncertainty in LLMs have been examined by Zhou et al. (2023) and Mielke et al. (2022), who looked at the impact of naturalistic expressions and linguistic calibration respectively. Their findings provide valuable insights for our exploration of multilingual LLMs.

**Closed-book QA** Question-answering (QA) traditionally relies on information retrieval and a model that generates answers from the retrieved text. Roberts et al. (2020) first proposed a *closed-book* QA approach, using only the question and the model's internal knowledge. They reported competitive performance against open-domain systems, however, later studies attributed these results to test-train overlap (Lewis et al., 2021), not inherent model knowledge. Wang et al. (2021) confirmed these findings, showing BART's (Lewis et al., 2020) limitations in memorisation and knowledge retrieval. Conversely, Brown et al. (2020) demonstrated GPT-3's high performance in factual QA. Peinl and Wirth (2023) built an English-only dataset (with questions on physics, math puzzles, metaphors, etc.) and focus on comparing the performance of a large number of language models. Recent studies have started to explore long-form QA (Amplayo et al., 2023), requiring more complex information use.

**Multilinguality** While the majority of language models are trained almost exclusively on English data, the exploration of their multilingual capabilities is an active field of research. Though their performance is often on par with existing state-of-the-art cross-lingual models and translation models (Winata et al., 2021), it increases further when explicitly trained on multilingual data (Lin et al., 2022). Shi et al. (2022) demonstrated the strong multilingual reasoning abilities of the models using chain-of-thought prompting on a mathematical task, even for low-resource languages. However, when Zhang et al. (2023) investigated similar behaviour, the model behaved in a translating manner, meaning performance can be affected by tasks where data cannot be explicitly translated. The closest work to ours is Sen et al. (2022) who built a multilingual QA dataset based on Wikipedia entities and with different question categories. They cover 9 different languages (4 of them overlap with ours) and presented baseline results including mT5 (Xue et al., 2021) language model and T5-CBQA (Roberts et al., 2020). We propose here an extension of an English dataset into a multilingual dataset, including 6 languages and present the performance of GPT-3.5, while also being interested in measuring the certainty when responding to such questions and how this relates to the change of accuracy across languages.

## 3 Methodology

### 3.1 Languages

Our analysis covers six languages with different amounts of resources available. Joshi et al. (2020) introduce a taxonomy to classify languages according to the amount of resources they have available on a scale of 0 (exceptionally limited resources) to 5 (quintessential rich-resource). Three of our analysed languages fall into the highest category (5): **English**, **German** and **Spanish**. **Dutch** and **Hindi** are categorised as Level 4: comparable to Level 5, but challenged by a lesser amount of labelled data. With **Amharic** we cover one Level 2 language, meaning a small amount of labelled datasets has been collected for it, but overall resources are very limited.

### 3.2 Dataset and task

We use a further modified version of the dataset used by Mielke et al. (2022). Theirs itself is a modified version of the TriviaQA dataset (Joshi et al., 2017). Originally a reading comprehension task, it features complex questions about trivia topics, that were intended to need cross-sentence reasoning to

|  | **What planet orbits between Uranus and Pluto?**<br>*Gold answer:* Neptune |
|---|---|
| en | There is no planet that orbits between Uranus and Pluto. |
| am | የትኛው ፕላኔት ነው የኢንተርኔት ፕላኔት (Internet Protocol) ነው። |
| de | Es gibt keinen Planeten, der zwischen Uranus und Pluto kreist. |
| es | El planeta que orbita entre Urano y Plutón es Neptuno. |
| hi | मंगल ग्रह परिक्रमा करता है। |
| nl | Er is geen planeet die tussen Uranus en Pluto draait. |

Example 1: Example answers where all, except Spanish are wrong and should have expressed uncertainty or lack of knowledge.

find answers.

To exclusively assess the knowledge encoded within the model's weights, Mielke et al. (2022) adapt the task to create a closed-book QA setting. All evidence documents associated with the questions were eliminated. They also removed the "disambiguation" suffix from the aliases originating from Wikipedia and employed these aliases to compile a set of permissible gold answers.

To create our dataset, we randomly extracted a subset of 1000 questions from the original training set utilised in their study. These questions were then translated into our target languages, specifically Amharic (am), German (de), Spanish (es), Hindi (hi), and Dutch (nl). English, German, Spanish and Dutch are written using the Latin script, while Amharic is written in Ge'ez script and Hindi in Devanagari script. For the translation process, we utilised the deep-translator library[2], which accessed Google Translate. The translations were validated during the subsequent annotation process. As a result, our closed-book QA dataset encompasses a total of 6000 question-answer pairs across the covered languages.

### 3.3 Model selection

For our analysis, we consider models that are capable of answering questions in a sentence format suitable for linguistic confidence analysis. Our initial consideration focuses on GPT-based models (`text-davinci-002`, `text-davinci-003`, `gpt-3.5-turbo`) (Brown et al., 2020), and publicly available open source models such as LLaMA (Touvron et al., 2023), BLOOM (Scao et al., 2023), and mT5 (Xue et al., 2021). We look at factors such as complexity of model setup, compute requirements, and language support in choosing which language model to investigate. Due to the computational cost involved in setting up publicly available base models, we instead opt for GPT-3.5[3]. All of our target languages except Amharic are explicitly listed as being part of GPT-3's training dataset[4]. Given that there is no transparent description of the full sources of GPT-3.5's training data, there is no guarantee that the model has not seen the language at all. Therefore we still decide to measure the performance of Amharic as zero-shot (or possibly few-shot) language probing. We specifically use the GPT3.5 variant `text-davinci-003` since it allows access to the log probabilities with its output. We require this score as a proxy to measure how confident the model is in its generated output sequence.

### 3.4 Annotation

The annotation scheme utilised in our study is derived from the methodology presented in Mielke et al. (2022). Each instance in the original dataset includes a question, a gold answer and a list of possible aliases to facilitate the annotator's job when assessing correctness (given that the generative model could realise a correct answer in different ways). During our annotation process, the first step was to validate that both the question and the correct gold answer were sensical and properly translated into each target language. Each question-answer pair was evaluated one by one by a native speaker, corrections were made to improve readability and keep the semantics of the original English question in case the annotator considered it adequate. As for the aliases, we left the original English and appended the translated aliases to the list of alternative correct answers to avoid confusions when e.g. translating Named Entities For example, if the original answer was the band *The Monkeys*, the correct answer should still be the original name of the band, not the translation *los monos*. Next, following the original guidelines, each automatically generated response was manually annotated according to the following criteria:

**Linguistic confidence** Labels are split into high (`HI`: confidently answers), low (`LO`: expresses uncertainty), and none (`DK`: admits not to know).

**Correctness** is divided fourfold, `Right` (correct answer and no incorrect additions), `Extra` (correct answer with added incorrect knowledge), `Wrong`

[2] https://github.com/nidhaloff/deep-translator

[3] For a detailed analysis of concerns involved in this choice, please see the Limitations section.

[4] github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv

(incorrect but not absurd answer), and Other (absurd/unrelated/no answer).

**Not classifiable** We add two labels to the Not classifiable class. The original had only OT (off-topic) in case the answer completely ignores the question. We add NI (not intelligible), the answer is not readable as a sentence e.g. due to incorrect syntax. Second, we add CO (Cut-off) for when the answer is too verbose and cut off before getting to the relevant part. For a fully annotated example see Appendix 8.

Due to the high cost associated with data collection, we only have one annotator per language, all have linguistic training and are both proficient in English and native speakers of their respective languages. Parallel annotation as such did not occur, however, all annotators were working with exactly the same questions (in their respective language), therefore a discussion was conducted among them regarding the edge cases, in order to share a similar decision-making process when judging them. Given that this dataset deals with factual answers, and that we followed a specific annotation process and strict guidelines, there was not much space for ambiguity: the annotator had to judge the answer based on the gold answers and possible aliases from the dataset. In case of doubt, annotators were also encouraged to search the web for confirming the correctness of each answer.

# 4 Results and Discussion

In this section, we report our findings on the performance of the model in terms of its accuracy, perceived confidence and their interplay. We also investigate difficulties overall and for specific languages. We look at 1000 samples for all languages.

## 4.1 Accuracy

As expected, English shows the highest correctness (83.05%) of answers, followed by German (77.65%). Interestingly, Spanish (69.99%) performs worse than Dutch (73.61%), even though it is the higher resource language. Model performance drops drastically for Hindi, at just 12.11%, and even more so for Amharic with nearly all answers being incorrect (only 0.53% correct).

We also look at automatic match-based evaluation, judging based on whether aliases appear in the generated response. Here we see high-performing languages lose performance, and low-performance ones gain, possibly influenced by keywords appearing in otherwise unintelligible answers.

| | **label** | **en** | **am** | **de** | **es** | **hi** | **nl** |
|---|---|---|---|---|---|---|---|
| detailed labels | Right | **82.25** | 0.53 | **77.35** | **68.77** | 12.56 | **73.00** |
| | Wrong | 16.85 | **88.73** | 19.24 | 24.92 | **76.93** | 23.86 |
| | Extra | 0.80 | - | 0.30 | 1.22 | 0.41 | 0.61 |
| | NI | 0.10 | - | - | 0.10 | 0.72 | - |
| | OT | - | 10.74 | - | - | 1.03 | 0.40 |
| | CO | - | - | 2.61 | 2.34 | 1.13 | 1.21 |
| | Other | - | - | 0.50 | 2.64 | 7.21 | 0.91 |
| binary | Correct | 83.05 | 0.53 | 77.65 | 69.99 | 12.98 | 73.61 |
| | Incorrect | 16.95 | 99.47 | 22.34 | 30.01 | 87.02 | 26.39 |
| match | Correct | ↓ 80.50 | ↑ 3.91 | ↓ 68.70 | ↓ 65.50 | ↑ 14.50 | ↓ 69.20 |
| | Incorrect | 19.50 | 96.09 | 31.30 | 34.50 | 85.50 | 30.80 |

Table 1: Normalised Correctness values. In case of no value, the label was not assigned. For binary labels, *Right* and *Extra* were merged into *Correct*, and the rest into *Incorrect*. For match-based labels, arrows indicate whether scores in- or decrease compared to human annotation.

## 4.2 Linguistic Confidence

Expressed confidence is generally high, with only a handful of answers indicating some level of uncertainty across all languages. Wrong answers can be seen as opportunities for the model to express uncertainty or admit lack of knowledge, however, Table 2 shows that, in most cases, such opportunities were missed. For example, the model is 99% confident in English while being 82.3% accurate, whereas it is also 99% in German while being only 77.4% accurate; one would expect at least a slight decrease in confidence for German outputs. An instance illustrating this can be found in Example 1. In this case, the model exhibits a high level of confidence in asserting the absence of a planet between Uranus and Pluto, despite the correct answer being Neptune. This indicates that the model needs further calibration and specific prompt engineering to express its confidence verbally.

| | **en** | **am** | **de** | **es** | **hi** | **nl** |
|---|---|---|---|---|---|---|
| HI | **99.00** | 27.75 | **99.00** | **97.87** | **100.00** | **99.29** |
| LO | 1.00 | - | 0.60 | - | - | 0.51 |
| NI | - | **72.25** | - | - | - | - |
| CO | - | - | 0.40 | 2.13 | - | - |
| OT | - | - | - | - | - | 0.20 |

Table 2: Normalised Linguistic Confidence values.

**Correctness of Confident Answers** Figure 1 shows the overlap of correct and high-confidence answers. Due to the high performance on languages like English, German or Dutch, the gap is smaller, but overall the results show a highly overconfident model, bettered only if performance catches up with its expressed confidence. Hindi shows a stark contrast in this regard, with 100% confident answers while only 12% are correct.

### 4.3 Correlations with Perplexity

Results revealed statistically significant correlations, ranging from weak to moderate, between perplexity and correctness across all languages except Hindi. Pearson correlation coefficients were computed for each language[5], yielding values ranging from r = 0.23, p < .001 (English) to r = 0.40, p < .001 (Amharic), suggesting potential for enhanced calibration. It is important to exercise caution when interpreting the results for Amharic due to the limited number of correct answers (n < 10).

Regarding the association between perplexity and confidence, the findings displayed variability. Notably, a significant yet weak correlation emerged solely in Spanish (r = 0.25, p = .001) and Dutch (r = 0.10, p = .001). Conversely, for English, German, Hindi, and Amharic, the p-values indicated the absence of a significant correlation. Thus, the relationship between perplexity and confidence appears less consistent across the languages tested.

### 4.4 Identification of Difficult Questions

We investigate patterns in the questions that the model tends to get correct/incorrect in each language. To do so, we compute n-grams, ranging from size 1 to 7, filtering out stop-words as well as instances with frequency less than 5 over the whole question set. Table A7 shows examples of 1-grams most related to correct and incorrect answers, per language respectively[6].

We observed certain patterns across languages. For instance, questions about "island" tend to be answered correctly in German, Spanish, and Dutch, while questions about "olympic" tend to receive incorrect answers in English, German, and Spanish. These patterns might indicate a knowledge gap within the model itself. However, we also found indications of language bias in knowledge access, as mentioned by Zhang et al. (2023). For example, the most relevant word for correct answers in English is "American," while in Dutch, the model often provides incorrect responses related to countries like "Russia" and "Britain."

| | **% of answers in English** | **% correct in English** | **% correct overall** |
|---|---|---|---|
| am | 19.84 | ↑ 2.00 | 0.53 |
| hi | 18.90 | ↑ 44.00 | 12.56 |

Table 3: Percentages of answers that were given in English instead of the language of the question and how many of those were correct in comparison with the overall accuracy.

### 4.5 Lower-resource settings

As stated earlier, both Amharic and Hindi showed notably poor performance. However, there were distinct differences: while Hindi responses were incorrect but coherent, most of the Amharic responses were incomprehensible. English was used for 18.9% and 19.8% of all answers for Hindi and Amharic, respectively (see Table 3. Notably, only 2% of these English responses were correct for Amharic, while for Hindi, this proportion spiked to 44% - increasing by 31% compared to overall accuracy. Interestingly, while Hindi and Dutch are in the same resource class according to Joshi et al. (2020), the accuracy of Hindi as a non-Latin script language is significantly lower.

For the case of Amharic, the model still generated some valid words, although the meaning of full sentences was non-sensical. The few cases that were judged as correct, were because the model generated the right (English) named entity as an answer. This could mean that the model saw at least some Amharic text during training, but given the extremely limited amount of data, the model cannot make much sense of the prompts that are being submitted.

## 5 Conclusion

Our work highlights performance disparities in LLMs across languages and shows a prevalent cross-language expressed overconfidence in responses in a closed-book setting. These findings underscore the importance of nuanced, multilingual calibration in LLMs. Accordingly, we hope to spur further progress by releasing our multilingual dataset and annotations.

[5] The Pearson correlation coefficients and p-values were calculated using scipy.stats.pearsonr from the SciPy library.

[6] Ngrams of size > 1 did not show any patterns

## Limitations

In our study, we acknowledge several limitations associated with the use of GPT-3.5 by OpenAI. Firstly, there is a cost factor involved in utilising this model due to its computational requirements and access restrictions. Additionally, the exact details of the data sources used in training GPT-3.5 are unknown, which can raise concerns about potential biases or inaccuracies in the model's responses. Similarly, the closed nature of GPT-3.5. as a proprietary model means the inner workings and specifics of its architecture and training process are not fully disclosed, limiting transparency and the ability for independent verification.

Furthermore, it is important to note that (`text-davinci-003`) is scheduled for deprecation in January 2024[7]. This time-frame imposes challenges in replicating and verifying the results of our study in the future. It was announced after the completion of this study.

Regarding the annotation process, we acknowledge that the limited number of annotators (only one) may pose a limitation. While we employ expert annotators, a larger pool of annotators would have been beneficial, as it could have provided a broader range of perspectives and potentially increased the reliability of the annotations.

## Acknowledgements

This research was partially funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) via the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen, the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022), Huawei Finland through the DreamsLab project, and by the EU Horizon 2020 project InTaVia: In/Tangible European Heritage - Visual Analysis, Curation and Communication (http://intavia.eu) under grant agreement No. 101004825.

## References

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query refinement prompts for closed-book long-form QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7997–8012, Toronto, Canada. Association for Computational Linguistics.

Lyle A Brenner, Derek J Koehler, Varda Liberman, and Amos Tversky. 1996. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, 65(3):212–219.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

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*, pages 1321–1330. PMLR.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-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, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training

[7] https://platform.openai.com/docs/deprecations

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Sarah Lichtenstein and Baruch Fischhoff. 1977. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20(2):159–183.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2022. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

René Peinl and Johannes Wirth. 2023. Evaluation of medium-large language models at zero-shot closed book generative question answering.

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)*, pages 5418–5426, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, and François Yvon et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *International Conference on Learning Representations (ICLR)*.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan L. Boyd-Graber. 2022. Re-Examining Calibration: The Case of Question Answering. In *Conference on Empirical Methods in Natural Language Processing*.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust gpt when your question is not in english.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models.

# A Appendix

## A.1 Average Perplexity

| | **en** | **am** | **de** | **es** | **hi** | **nl** |
|---|---|---|---|---|---|---|
| HI | **0.87** | 0.78 | **0.89** | **0.88** | **0.88** | **0.88** |
| LO | 0.81 | - | 0.82 | - | - | 0.84 |
| NI | - | **0.87** | - | - | - | - |
| CO | - | - | **0.89** | 0.86 | - | - |
| OT | - | - | - | - | - | 0.85 |

Table A 4: Average perplexity per confidence label

| | **en** | **am** | **de** | **es** | **hi** | **nl** |
|---|---|---|---|---|---|---|
| RIGHT | **0.88** | **0.88** | **0.90** | **0.89** | **0.88** | **0.89** |
| WRONG | 0.82 | 0.86 | 0.86 | 0.86 | **0.88** | 0.84 |
| EXTRA | 0.84 | - | 0.87 | 0.84 | 0.87 | 0.90 |
| NI | 0.75 | - | - | 0.86 | 0.91 | - |
| OT | - | 0.69 | - | - | 0.78 | 0.83 |
| CO | - | - | 0.88 | 0.87 | 0.87 | 0.86 |
| OTHER | - | - | 0.86 | 0.84 | **0.87** | 0.84 |

Table A 5: Average perplexity per correctness label

**Difficulty with Wh-Questions** We investigate the performance, according to the binary correctness labels, for different Wh-question, for each language, and across languages, as shown in Table 6. For this, we first search for Wh keywords in the English corpus, and select the corresponding translated questions.

The highest performance is achieved by "What" and "How" questions, while "Who", "Where" and "Which" have similar performances[8].

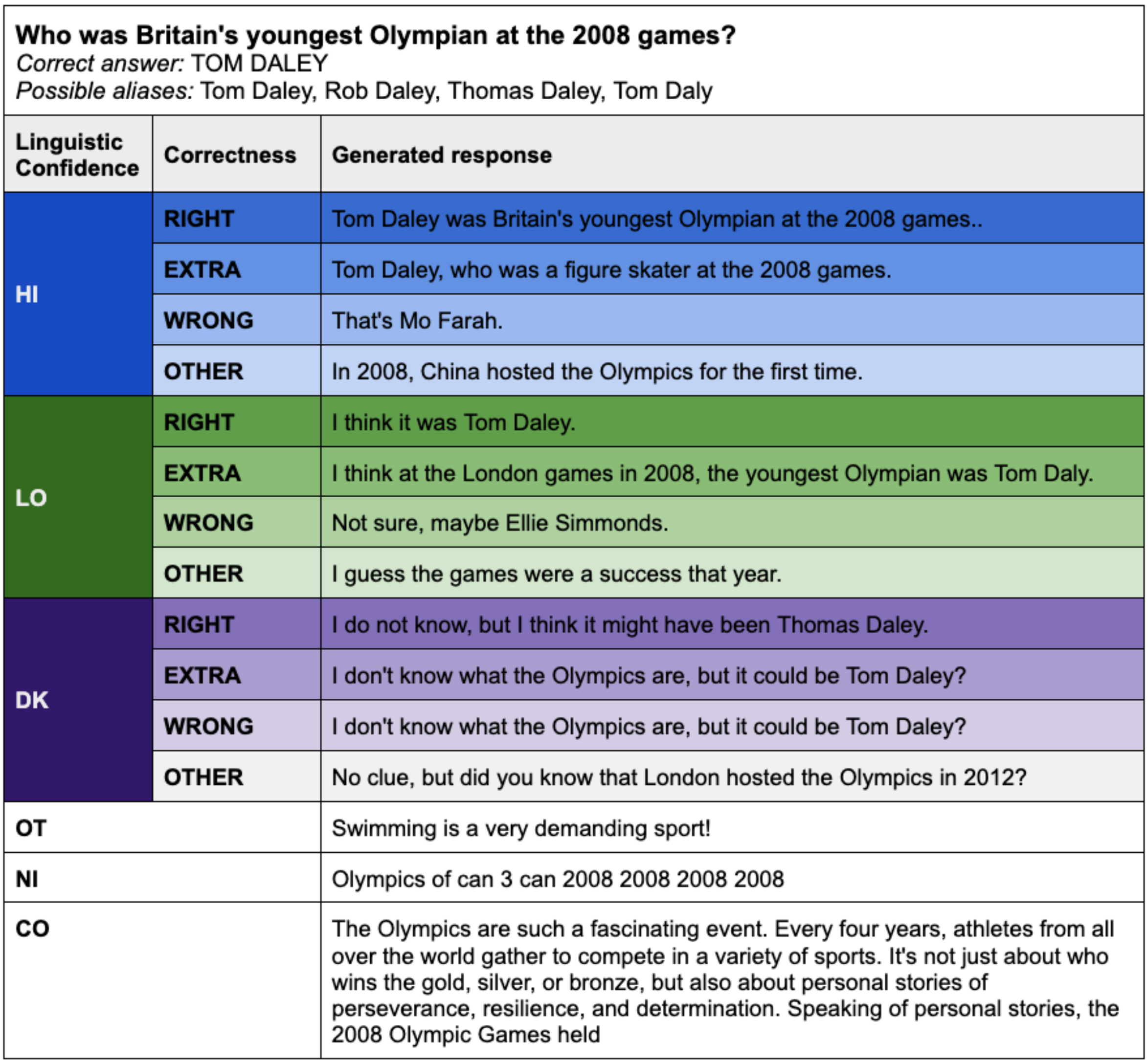
| | **en** | **am** | **de** | **es** | **hi** | **nl** | total |
|---|---|---|---|---|---|---|---|
| Who (110) | 84.5 | 1.0 | 84.4 | 70 | 12.0 | 74.3 | 54.9 |
| What (219) | 85.3 | 0.5 | 84.1 | 72.8 | 13.5 | 76.0 | 57.2 |
| Where (8) | 87.5 | 0.0 | 75 | 71.4 | 0.0 | 75.0 | 51.1 |
| When (1) | 0.0 | 0.0 | 0.0 | 0.0 | - | 0.0 | 0.0 |
| Which (193) | 78.7 | 0.7 | 75.9 | 66.3 | 11.9 | 71.5 | 53.2 |
| How (27) | 81.4 | 0.0 | 66.6 | 80 | 20 | 70.3 | 55.8 |

Table A 6: Accuracy on "Correct" class, per type of WH-question, Sample size in parenthesis.

[8] We exclude "When" due to the low sample size

| lang | 1-gram | %correct | %incorrect | 1-gram | %correct | %incorrect |
|---|---|---|---|---|---|---|
| en | american | 96.0 | 4.0 | years | 42.9 | 57.1 |
|  | known | 95.0 | 5.0 | half | 50.0 | 50.0 |
|  | capital | 94.7 | 5.3 | olympic | 50.0 | 50.0 |
| am | የሚታወቀው (known) | 16.7 | 83.3 | የትኛው (which one) | 1.5 | 98.5 |
|  | በፊት (past) | 16.7 | 83.3 | ከተማ (city) | 2.2 | 97.8 |
|  | የነበረው (which was) | 16.7 | 83.3 | ዋና (main) | 3.6 | 96.4 |
| de | gegründet (founded) | 93.8 | 6.2 | olympischen (olympic) | 25.0 | 75.0 |
|  | insel (island) | 92.3 | 7.7 | einzigen (single) | 33.3 | 66.7 |
|  | rolle (role) | 91.7 | 8.3 | vereinigten (united) | 42.9 | 57.1 |
| es | escribió (wrote) | 95.7 | 4.3 | olímpico (olympic) | 16.7 | 83.3 |
|  | principal (main) | 94.1 | 5.9 | actual (current) | 33.3 | 66.7 |
|  | isla (island) | 93.3 | 6.7 | tres (three) | 33.3 | 66.7 |
| hi | मृत्यु (death) | 50.0 | 50.0 | सबसे (most) | 3.8 | 96.2 |
|  | बार (times) | 42.9 | 57.1 | शहर (city) | 4.2 | 95.8 |
|  | लिखा (written) | 42.9 | 57.1 | भूमिका (role) | 7.7 | 92.3 |
| nl | speelde (played) | 95.4 | 4.6 | verscheen (appeared) | 28.6 | 71.4 |
|  | hoofdstad (capital) | 94.4 | 5.6 | rusland (Russia) | 33.3 | 66.7 |
|  | eiland (island) | 92.9 | 7.1 | brittannië (Britain) | 42.9 | 57.1 |

Table A 7: Top 1-grams most related to correct (left) and incorrect (right) answers, per language respectively. English translations in parenthesis.



**Who was Britain's youngest Olympian at the 2008 games?**
*Correct answer:* TOM DALEY
*Possible aliases:* Tom Daley, Rob Daley, Thomas Daley, Tom Daly

| Linguistic Confidence | Correctness | Generated response |
|---|---|---|
| HI | RIGHT | Tom Daley was Britain's youngest Olympian at the 2008 games.. |
|  | EXTRA | Tom Daley, who was a figure skater at the 2008 games. |
|  | WRONG | That's Mo Farah. |
|  | OTHER | In 2008, China hosted the Olympics for the first time. |
| LO | RIGHT | I think it was Tom Daley. |
|  | EXTRA | I think at the London games in 2008, the youngest Olympian was Tom Daly. |
|  | WRONG | Not sure, maybe Ellie Simmonds. |
|  | OTHER | I guess the games were a success that year. |
| DK | RIGHT | I do not know, but I think it might have been Thomas Daley. |
|  | EXTRA | I don't know what the Olympics are, but it could be Tom Daley? |
|  | WRONG | I don't know what the Olympics are, but it could be Tom Daley? |
|  | OTHER | No clue, but did you know that London hosted the Olympics in 2012? |
| OT |  | Swimming is a very demanding sport! |
| NI |  | Olympics of can 3 can 2008 2008 2008 2008 |
| CO |  | The Olympics are such a fascinating event. Every four years, athletes from all over the world gather to compete in a variety of sports. It's not just about who wins the gold, silver, or bronze, but also about personal stories of perseverance, resilience, and determination. Speaking of personal stories, the 2008 Olympic Games held |

Table A 8: Examples of all categories in the annotation scheme.