LUCY: *L*inking *U*ncertainty and *C*onsistenc*Y* of Large Language Models for Question Answering

Urja Khurana
Vrije Universiteit Amsterdam
u.khurana@vu.nl

Lea Krause
Vrije Universiteit Amsterdam
l.krause@vu.nl

1 Introduction

The increasing deployment of large language models (LLMs) in real-world applications highlights the need to understand their reliability and generalization capabilities. While being robust is crucial, i.e. generalizing to new data but the same task (Hupkes et al., 2023), it is equally important that models are consistent in their generalization across different runs. Previous studies have explored consistency (Jang et al., 2022; Bartsch et al., 2023; Madaan et al., 2024; Weber et al., 2023; Elazar et al., 2021; Khurana et al., 2021) but have largely overlooked its relationship with uncertainty. This paper examines the link between consistency and uncertainty in LLMs. We hypothesize that uncertain models are less consistent across multiple runs. To test this, we analyze the behavior of several LLMs on question-answering tasks, both in open and closed-book settings, using metrics like log-likelihood (for uncertainty) and Fleiss' Kappa (for consistency). We conduct experiments across five random seeds on four English and four multilingual datasets to assess the robustness of these models.

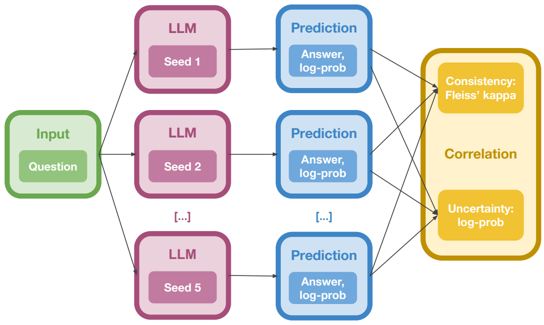

Figure 1: Experimental setup: For each input question, answers are generated for several seeds. The model's consistency is then correlated with the sequence probability of the answer.

2 Methodology

2.1 Datasets

We selected question-answering (QA) datasets to examine the link between consistency and uncertainty in large language models, as they require precise answers, facilitating consistency assessment across runs. We included both English-only and multilingual datasets:

English datasets: *TruthfulQA* (Lin et al., 2022), a closed-book set designed to test truthfulness and expression of uncertainty; *CoQA* (Reddy et al., 2019), a conversational QA dataset evaluated on the first question in each dialogue; *TriviaQA* (Joshi et al., 2017), in a closed-book format; and *SQuAD 1.1* (Rajpurkar et al., 2016), aligned with XQUAD English for consistency with multilingual evaluations.

For multilingual analysis, we used *xTriviaQA* (Krause et al., 2023), a translated version of TriviaQA in five languages; *TyDiQA* (Clark et al., 2020), an open-book dataset covering typologically diverse languages in the gold-passage setting; and *XQuAD* (Artetxe et al., 2020), a cross-lingual subset of SQuAD, translated into 12 languages.

2.2 Models

We selected four widely-used large language models for our experiments: OPT (Zhang et al., 2022), LLAMA-2 (Touvron et al., 2023), BLOOM (Scao et al.), and GPT-4 (OpenAI et al., 2024). We chose model variants with parameter counts close to 7 billion for comparability, except for GPT-4, where the parameter count is undisclosed. We used the *chat* version of LLAMA-2 and the *GPT-4o mini* variant for GPT-4.

2.3 Metrics

To evaluate model performance and behavior, we employed metrics that capture robustness, uncertainty, and correctness. These metrics provide insights into model consistency and confidence across different conditions.

2.3.1 Robustness

We assessed robustness by measuring consistency in responses across different seeds within the same model family, using two metrics: *Fleiss' Kappa* and *Model Disagreement Variation*.

Fleiss' Kappa quantifies agreement between annotators, ranging from 0 (no agreement) to 1 (perfect agreement). In our context, different seeds of the same model family act as annotators, and predictions are treated as annotations. We use BERTScore with a threshold of 0.8 to determine if two answers are the same.

Model Disagreement Variation examines agreement on the presence of the correct answer across models, following the approach by Mostafazadeh Davani et al. (2022). A score of 0 indicates full agreement, while 1 indicates no agreement.

2.3.2 Uncertainty

We estimate uncertainty in sequence-prediction tasks by calculating the log-probability of the sequence. We use the geometric mean, which as discussed by Malinin and Gales (2022), is sensitive to low-probability events, providing a normalized certainty measure across sequences.

2.3.3 Correctness

We evaluate correctness using three metrics: ROUGE-L (Lin, 2004), BERTScore (Zhang* et al., 2019), and a Presence Metric, which checks for the occurrence of reference answers in model predictions, providing a basic measure of correctness.

3 Experimental Setup

Figure 1 outlines our experimental setup. For each dataset, we perform inference over the validation set five times with different seeds. The models are prompted with the *context* (for open-book datasets) and the *question*. Since our focus is on the relationship between consistency and uncertainty, optimizing prompts is out of scope.

We use HuggingFace `transformers` for inference with quantization on all models except GPT-4, where we rely on the `OpenAI API`. A global seed is set for the `transformers` models and the corresponding parameter for `OpenAI API`. Responses are capped at 40 tokens, and we set *top_p* to 0.95 for `transformers`.

4 Results

Figures 2 and 3 present our results for LLAMA-2, showing the relationships between Fleiss' Kappa and sequence log-likelihood, as well as Model Disagreement (MD) and sequence log-likelihood. The Pearson and Spearman correlations, shown in the plots, confirm these trends. We see a positive correlation between Fleiss' Kappa and sequence log-likelihood, while the relationship between MD and sequence log-likelihood is negative.

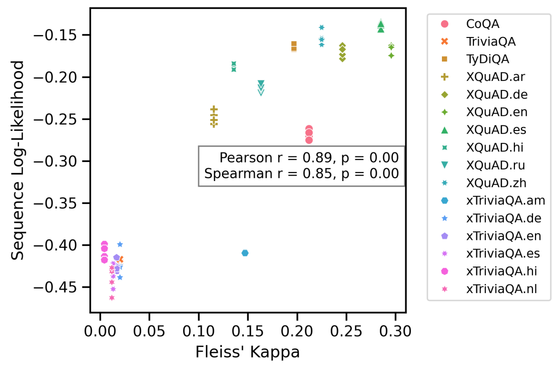

Figure 2: Fleiss' Kappa vs. Sequence Log-Likelihood for LLAMA-2.

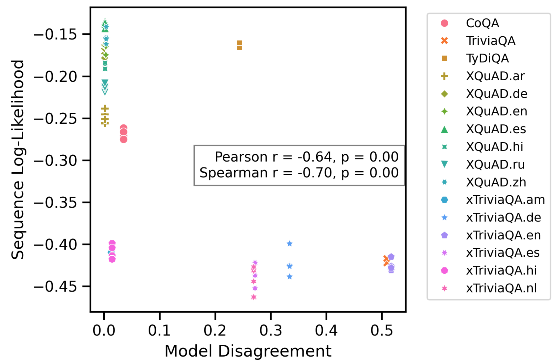

Figure 3: Model Disagreement vs. Sequence Log-Likelihood for LLAMA-2.

5 Conclusion

The initial results from our experiments are promising, showing a significant correlation between model robustness and uncertainty across different seeds: lower uncertainty indicates higher consistency. However, this correlation does not extend to correctness metrics. Given that our open-ended QA implementation may introduce unforeseen factors, such as tokenization issues in underrepresented languages, we plan to extend our analysis to include Multiple-choice QA for a more controlled evaluation.

Acknowledgments

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022), and the Spinoza grant (SPI 63-260) awarded to Piek Vossen.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 89–105, Singapore. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni, editors. 2023. *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*. Association for Computational Linguistics, Singapore.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2021. How emotionally stable is ALBERT? testing robustness with stochastic weight averaging on a sentiment analysis task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lea Krause, Wondimagegnhue Tufa, Selene Baez Santamaria, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: Exploring the calibration and expression of (un)certainty of large language models in a multilingual setting. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 1–9, Prague, Czech Republic. Association for Computational Linguistics.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*.

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

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and Niklas Muennighoff. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel

Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *Preprint*, arXiv:2205.01068.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.