Title: **Who What When Did: Multiple WH-questions in Large Language Models**
Author: Mădălina Zgreabăn

Abstract

Good understanding of questions and answers improves the performance and user satisfaction (Bender et al., 2021) of large language models (LLMs), as well as their diversity, through rare phenomena such as multiple WH-questions (MWHqs). However, despite these benefits, questions in general, and more complex phenomena, such as MWHqs, have been neglected, especially in multilingual models (Ruder and Sil, 2021).

For example, cross-linguistically, MWHqs are very complex, being ungrammatical (e.g. Italian), permitted in-situ (e.g. English), or fronted (e.g. Romanian), while their answers can be predominantly *mention-all* and *mention-some*, such as in Hindi (Boškovic, 1998) or Romanian, or exclusively *mention-all* answers, such as German (Foryś-Nogala et al., 2017). Semantically, semantic quantifiers correlates with ability to provide mention-all answers (Foryś-Nogala et al., 2017).

To account for these research gaps, the current research proposal proposes firstly to test if LLMs are cross-linguistically sensitive to the (un)grammaticality of MWHqs, and then if training models on quantifiers increases their ability to provide *mention-all* answers when compared to normal LLMs.

Thus, the main research questions are 'What are the semantic abilities of LLMs in WH-questions and, if any, how human-like are they?' with the following research sub-questions:

sRQ1: Are multilingual LLMs sensitive to MWHqs?
sRQ2: Do LLMs expect more *mention-all* or *mention-some* answers depending on the exhaustivity of the question?
sRQ3: Are LLMs fine-tuned on structures correlated with improved exhaustivity more sensitive to *mention-all* or *mention-some* answers?

In the first experiment, two datasets of fronted and in-situ MWHqs will be created for Italian, English and Romanian, used to estimate the surprisal of LLMs. Given their ungrammaticality, we expect generally largest surprisal scores for MWHqs for Italian, as well as bigger surprisal scores for fronted rather than in-situ MWHqs for English, under the hypothesis that LLMs have some cross-linguistic information about the phenomenon. No difference in surprisal scores is expected between MWHqs and fronted MWHqs in Romanian, given both structures are grammatical. Contrastively, no surprisal scores between any categories across languages would be expected, if LLMs would not be sensitive to such cross-linguistic differences.

In the second experiment, we will train models on sentences with more quantifiers for English and Romanian. Generally, under the hypothesis that models have semantic knowledge about answers, we expect bigger averaged surprisal scores for *mention-some* answers given to exhaustive questions than for *mention-all* answers, a difference we do not expect for non-exhaustive questions. No difference in surprisal values is expected under the hypothesis the models do not have semantic knowledge about questions. We expect the trained models to have bigger surprisal for *mention-some* answers to exhaustive questions, under the hypothesis models learn human language cues (i.e. quantifiers), in line with Frank et al. (2015) or Michaelov et al. (2023). No such difference is expected if the learned cues are not human-like.

Finally, this research proposal would offer insights into the diversity of NLP tools, while raising awareness about the current semantic abilities of LLMs. The results could be further compared to those of human experiments.

References

Achiam, O.J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., … & Zoph, B. (2023). GPT-4 Technical Report.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Boškovic, Z. (1998). On the interpretation of multiple questions.

Foryś-Nogala, M., Haman, E., Katsos, N., Krajewski, G., & Schulz, P. (2017). Syntactic, semantic and pragmatic correlates of the acquisition of exhaustivity in wh-questions: A study of Polish monolingual children. Language Acquisition, 24(1), 27-51.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998-998.

Gilbert, H., Sandborn, M., Schmidt, D. C., Spencer-Smith, J., & White, J. (2023, November). Semantic compression with large language models. In 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-8). IEEE.

Kotek, H. (2016). On the semantics of wh-questions. In Proceedings of Sinn und Bedeutung (Vol. 20, pp. 430-447).

Lam, S. Y., Zeng, Q., Zhang, K., You, C., & Voigt, R. (2023). Large Language Models Are Partially Primed in Pronoun Interpretation. arXiv preprint arXiv:2305.16917.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong Prediction: Language model surprisal explains multiple N400 effects. Neurobiology of Language, 1-71.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., ... & Raffel, C. (2022). Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.

Roelofsen, F., & Dotlačil, J. (2023). Wh-questions in dynamic inquisitive semantics. Theoretical Linguistics, 49(1-2), 1-91.

Ruder, S., & Sil, A. (2021, November). Multi-domain multilingual question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts (pp. 17-21).

OpenAI. 2021. ChatGPT 3.5. OpenAI Blog. https://openai.com/blog/chatgpt.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Saba, W. S. (2023, October). Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In International Conference on Conceptual Modeling (pp. 3-19). Cham: Springer Nature Switzerland.

Sinha, K., Parthasarathi, P., Pineau, J., & Williams, A. (2020). Unnatural language inference. arXiv preprint arXiv:2101.00010.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. Cerebral Cortex, 26(6), 2506-2516.

Willis, P. M. (2008). The role of topic-hood in multiple-wh question semantics. In Proceedings of the 27th West Coast Conference on Formal Linguistics Poster Session (pp. 87-95).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.