Evaluating Hallucinations in Large Language Models for Bulgarian Language

Melania Berbatova
Sofia University 'St. Kliment Ohridski'
msberbatova@fmi.uni-sofia.bg

Yoan Salambashev
Sofia University 'St. Kliment Ohridski'
jsalambash@uni-sofia.bg

Abstract

In this short paper, we introduce the task of evaluating the hallucination of large language models for the Bulgarian language. We first give definitions of what is a hallucination in large language models and what evaluation methods for measuring hallucinations exist. Next, we give an overview of the multilingual evaluation of the latest large language models, focusing on the evaluation of the performance in Bulgarian on tasks, related to hallucination. We then present a method to evaluate the level of hallucination in a given language with no reference data, and provide some initial experiments with this method in Bulgarian. Finally, we provide directions for future research on the topic.

1 Introduction

Hallucination in large language models (LLMs) refers to the generation of non-factual statements or information that cannot be verified from the source. The latest generative language models, such as Llama, GPT-4 and other GPT-based models, are known to suffer from hallucination problems. The lack of trustworthiness of the generated outputs of LLMs is one of the main factors that stop their employment in sectors like education and healthcare, where there are high standards for factual accuracy. While numerous annotated evaluation datasets and benchmarks for evaluating the level of hallucinations exist for the English language, it is not the same for most human languages. Evaluation of hallucination in lower-resource languages[1], such as Bulgarian, is still an open research problem.

Due to the lack of annotated data on hallucinations in Bulgarian, we chose to work with a Zero-resource evaluation method called *SelfCheckGPT* (Hardalov et al., 2020), which offers an approximate estimation of the amount of hallucinations in the text. We experimented with data from Bulgarian matriculation exams, part of the EXAMS dataset (Hardalov et al., 2020), which we processed to derive prompts for text generation on different school subjects.

2 Definitions

2.0.1 What is a Hallucination?

Hallucination in LLMs is still an open research problem and there is not a universal definition of the term. According to Ji et al. (2023), there exist two categories of hallucination: intrinsic and extrinsic. *Intrinsic hallucinations* refer to the model's generated text that contradicts the source or input. Cases where intrinsic hallucinations occur are summarization, machine translation and other tasks in which and input text is given. *Extrinsic hallucinations* refer to the model's generations that cannot be verified from the source/input content (or in other words, output that can neither be supported nor contradicted by the source). Extrinsic hallucinations can occur in all text generation tasks. (Bang et al., 2023)

Other authors, such as Preetham, add other types of hallucinations, like *nonsensical statements*, where the model generates a response that doesn't make sense or is unrelated to the context, and *improbable scenarios*, where the model generates a response that describes an implausible or highly unlikely event. Hallucination in large language models can also be related to the model's inability to produce factual and commonsense knowledge (F. Petroni and Riedel, 2019) or low degree of truthfulness, the measure of whether a language model is truthful in generating answers to questions (Lin et al., 2022).

2.1 Evaluation Methods

We group the observed hallucination methods into three main groups: fact-checking evaluation, hu-

[1] We use the term ”lower-resource language” instead of ”low-resource”, as Bulgarian is sometimes referred as ”low-resource” and other times as ”medium-resource”, depending on the definitions different authors use.

Proceedings of the Student Research Workshop associated with RANLP-2023, pages 55–63, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_006

man evaluation, and counterfactual evaluation, proposed by (Preetham, 2023).

1. *Fact-checking evaluation* (F. Petroni and Riedel, 2019; Kassner et al., 2021; Jifan Yu, 2023) involves comparing the generated outputs of a model with a knowledge base or trusted sources to ensure that the facts presented in the generated text are accurate and supported by evidence.

2. *Human evaluation* (Lin et al., 2022; Li et al., 2023; Manakul et al., 2023) involves employing human evaluators to assess the relevance and truthfulness of the generated outputs. This evaluation metric leverages human judgment to provide insights into the subjective aspects of generated outputs.

3. *Contrastive Evaluation* (Manakul et al., 2023) involves presenting the model with a set of alternative completions or responses, where some options may include hallucinated information. This metric evaluates the model's ability to select the correct or most plausible output among the alternatives.

3 Related Work

3.1 Multilingual Evaluation

There are numerous publications on the performance of large language models in multilingual settings, both provided by the researchers developing large multilingual language models, or independent research groups. In this section, we would focus on the multilingual evaluation of the latest large generative language models, relevant to the Bulgarian language. Previous multilingual large language models, such as mBERT (Devlin et al., 2018), mBART (Liu et al., 2020) and mT5 (Xue et al., 2020) would stay outside the scope of the current research.

One of the first attempts towards a multilingual GPT-based model is XGLM (Lin et al., 2021), based on the GPT-3 architecture but trained on more than 100 languages, including Bulgarian. Lin et al. evaluated XGLM on the XNLI dataset (Conneau et al., 2018) for natural language inference and found that for the Bulgarian language, multilingual training significantly improves the results compared to monolingual training in GPT-3, but still lags behind the results of the combination of monolingual training and machine translation. Another interesting finding shared by XGLM's authors is that while most cross-lingual few-shot settings significantly improve over the 0-shot setting for the target language, Bulgarian is an exception, as it does not benefit from Russian, despite being in the same language family.

mGPT (Shliazhko et al., 2022) is another multilingual model, based on the GPT architecture. mGPT is trained on 61 languages from 25 language families and aimed at improving language understanding for the official and minor languages in Russia and former USSR countries. Authors also provide an interactive API of the model via the Hugging Face platform[2]. The model is evaluated on two tasks – language perplexity and knowledge probing. For the tasks of knowledge probing, which is a form of fact-checking evaluation of the ability of the language models to produce factual knowledge, they use the mLAMA probe (Kassner et al., 2021), which extends the original LAMA probe (F. Petroni and Riedel, 2019) to the multilingual setting. On this task, the performance of Bulgarian is lower than the average, meaning that the model fails at generating factual text in Bulgarian. This result aligns with our observations that the model often hallucinates, producing extrinsic hallucinations and nonsensical statements, when prompted in Bulgarian language, as shown in Table 1.

Recently, Ahuja et al. (2023) and Bang et al. (2023) perform multilingual analysis on the latest large language models, and while both works conduct a massive study on different languages, evaluation for the Bulgarian language is not present in either of them. However, they provide some valuable insights for lower-resource and non-Latin languages. Bang et al. state that despite ChatGPT's strong performance in many high-resource and medium-resource languages, the model still has problems in translating and generating text in languages that do not use the Latin script, even though these languages are considered high-resource. Moreover, Ahuja et al. suggests that one of the factors that lead to a decrease in performance in non-Latin languages is the fact that LLMs by default use a tokenizer build for the English language, which leads to incorrect tokenization of words in other languages. We found evidence of both claims in our experiments with ChatGPT, as the model sometimes responds in Russian, while prompted

[2] https://huggingface.co/ai-forever/mGPT

Extrinsic Hallucination	Nonsensical Statement
Столицата на България е най-големият град в Европа, а в него живеят над 1.5 милиона души.	Българите са най-бедни в Европа, но са най-бедни в света.

Table 1: Examples of mGPT text generation for Bulgarian language. Text in black is the prompt, and text in blue – the model generated text. English translations are shown in Table 6.

in Bulgarian, and sometimes generates text with words that are non-existent in Bulgarian, but resemble a truncated version of existing words. Finally, the authors of MEGA (Ahuja et al., 2023) also state that comprehensive assessment of LLMs for non-English languages is very challenging due to the scarcity of datasets available, which also motivated us to search for alternative approaches for the evaluation of hallucinations.

3.2 Zero-Resource Evaluation

When no reference data is present, the level of hallucination of generative models can be estimated in a zero-resource manner. This method is especially useful for lower-resource languages, for which annotated datasets and other publically available language resources are scarce. Manakul et al. (2023) propose the SelfCheckGPT method, which is a simple sampling-based approach that can predict whether responses generated by large language models are hallucinated or factual.

The underlying idea behind SelfCheckGPT is that when a large language model has a deep understanding of a specific concept, the responses it generates will tend to be similar and consistently contain factual information. Conversely, when the model generates hallucinated facts, the sampled responses are likely to diverge and may even contradict one another. By obtaining multiple responses through stochastic sampling from the LLM, it becomes possible to assess the level of information consistency among these responses. This approach enables the identification of factual statements versus those that are likely to be hallucinated, without relying on an external knowledge base.

4 Experiments

4.1 Experimental Setup

We decided to test the models in a black-box, zero-resource manner with the *SelfCheckGPT* framework, proposed by Manakul et al. (2023). The method proposes several evaluation scores, of which we chose Unigram and BERTScore, as they were most suitable for our experimental setup.

In order to create model-generated passages, suitable for black-box, zero-resource evaluation of hallucination, we use the EXAMS dataset (Hardalov et al., 2020), part of the bgGLUE (Hardalov et al., 2023) benchmark. It contains multiple choice questions from the Bulgarian marticulation exam in 6 subjects: Biology, Philosophy, Geography, History, Physics, and Chemistry.

In order to prepare the LLM prompts, we performed the following 3 steps:

1. Filter the dataset to preserve only the Bulgarian data.

2. Remove the irrelevant and non-informative items with no context/value in the actual question like 'Which statement is true for endocytosis?'. This way, our prompts are the open-ended questions (not having '?' in last/penultimate position).

3. Add a navigating prefix to the prompt for each question, "Напиши абзац, започващ с 'Q' translated as "Write a paragraph, starting with 'Q'", where Q is the question.

Input (question): Кондензатор със заряд q = 0,2 C и напрежение U = 4 V, има капацитет C равен на:

Output (prompt): Напиши абзац, започващ с 'Кондензатор със заряд q = 0,2 C и напрежение U = 4 V, има капацитет C равен на:'

As a result, we ended up with a total of 566 prompt questions. They are nearly equally distributed subject-wise: 130 in Biology, 136 in Philosophy, 75 in Geography, 87 in History, 70 in Physics and 68 in Chemistry. We ran 5 iterations of each prompt – the first one was used to generate the main passage for the evaluation, and the rest for the sampled passages.

4.2 Models

We chose to use the following LLMs:

- *text-davinci-003* by OpenAI
- *gpt-3.5-turbo-0613* (the model behind ChatGPT) by OpenAI (Brown et al., 2020)

Our decision to choose these models was based on the fact that they are two of the biggest (in terms of parameters) state-of-the-art large language models which are trained on Bulgarian language.

All experiments were run on both the OpenAI DaVinci and GPT-3.5 Turbo models. We generated the prompt responses using the OpenAI Completions API (model: text-davinci-003) and the Chat Completions API (model: gpt-3.5-turbo-0613) with a token limit of 300.

4.3 Evaluation

We evaluated for the factuality of the generated passages using (i) the BERTScore, (ii) average unigram, and (iii) maximum unigram scores, described in Manakul et al. (2023).

SelfCheckGPT with BERTScore finds the averages BERTScore of a sentence with the most similar sentence of each drawn sample. This method lies on the assumption that if the information in a sentence appears in many drawn samples, it is very likely that the information is factual, whereas if the statement appears in no other sample, it is more likely to be a hallucination. At the other hand, the unigram-based scores aim at approximating the original LLM's. The assumption of this method is that given the sample responses, one could train a new language model, which token probabilities would approximate the ones from the original LLM. [3]

BERTScore scores are in the interval [0.0, 1.0], and higher value estimates a higher chance of hallucination. Unigram scores are in the interval [0.0, +inf) and again a high value means a higher chance of hallucination.

We compute the BERTScore for each subject individually by calculating the average value of the relevant scores for each passage and then calculating the average of all those passage scores. Unigram scores are calculated by taking the average of all document-level scores per subject.

As the unigram scores diverge to infinity for some passages, we were forced to replace those values before the computation of the overall average score per subject. We decided that the most reasonable value substitute would be the maximum among the remaining values for each log probability score, respectively. The final evaluation results are shown in Table 2.

4.4 Results

In our evaluation of hallucination tendencies in LLMs in Bulgarian language, we examined two models: *text-davinci-003* and *gpt-3.5-turbo-0613*. Considering all the metrics we evaluated, the second model tends to hallucinate more. Philosophy has the highest evaluation score for most of the metrics, as the Philosophy questions were relatively broad (such as "Философията е..."(*"Philosophy is... "*)) and therefore resulted in more varying responses, compared to the ones for the rest of the subjects. We still lack a similar assessment of hallucinations in other languages, but the listed unigram scores are significantly higher than the ones shown in the SelfCheckGPT repository[4]. The referred BERTScores, however, are higher, as the authors decided to demonstrate the method with sentences that were quite different from each other. The average scores are summarized in Table 3.

We observe one specific type of hallucination that often occurs in the responses that we can conditionally call *"foreign language hallucination"*, which cover different kinds of language-specific errors, such as spelling errors, wrong word order, and misused words and phrases. What separate them from other nonsensical statements is that they make sense once the text is translated via a machine translation tool, such as Google Translate. An example of such case is the word "резониране" (rezonirane, "resonance"), used in the meaning of "reasoning" in the second example in Table 4. Even though "резониране" sounds similar to the English "reasoning", but does not exist in Bulgarian language with this meaning. Explanations of different kinds of *"foreign language hallucinations"* can be seen in Table 5.

One additional observation we made while conducting our experiments is that the sentence splitting function used in the SelfCheckGPT code does not perform well in Bulgarian language and therefore degrades the reliability of the assessment. In the future, we plan to change it to a sentence splitter

[3] Formulas and more detailed explanations can be found in the original paper.

[4] https://github.com/potsawee/selfcheckgpt

	text-davinci-003			gpt-3.5-turbo-0613		
Subject	**Avg-uni**	**Max-uni**	**BERTScore**	**Avg-uni**	**Max-uni**	**BERTScore**
Biology	4.2436	5.1093	0.0831	4.6059	5.9698	0.4496
Philosophy	**4.2849**	**5.1740**	0.0850	**4.6151**	**6.0434**	**0.5040**
Geography	4.1471	5.0612	**0.0906**	4.6097	5.9266	0.4635
History	4.2592	5.0796	0.0864	4.6096	5.8535	0.4780
Physics	4.0798	5.0322	0.0776	4.5365	5.8853	0.4765
Chemistry	4.1946	5.0716	0.0778	4.5009	5.8124	0.4598
Average	4.2170	5.0999	0.0837	4.5988	5.9431	0.4734

Table 2: Average evaluation scores for each subject with *text-davinci-003* and *gpt-3.5-turbo-0613*. Avg-uni means the Average unigram score and Max-uni – Maximum unigram score.

Avg-uni	**Max-uni**	**BERTScore**
3.2186	4.0254	0.2627

Table 3: Average evaluation scores from the experiments, provided in the SelfCheckGPT repository.

developed specifically for the Bulgarian language, proposed by Berbatova and Ivanov (2023).

5 Conclusion

In this short paper, we demonstrate our work in progress on the task of evaluating the level of hallucination of large language models in Bulgarian language. We give definitions of different types of hallucinations and methods for evaluation, make an overview of the related work, and provide some initial experiments.

Our research is aimed towards more equally spread employment of the latest technology across different languages. Researchers working on other lower-resource languages can use our work as a source of ideas generation and inspiration.

6 Future Work

In the future, we would like to extend our research in the following directions:

1. Further research on methods and datasets for hallucination evaluation. We would like to do a comprehensive overview of the latest benchmarks for hallucination evaluation, such as TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023), Kola (Jifan Yu, 2023) others, and analyze if a similar benchmark can be developed for Bulgarian.

2. Extend our experiments, so we have a more objective estimation on the level of hallucination of different LLMs in Bulgarian language. One of the first experiments that we want to conduct is to evaluate the latest LLMs, such as GPT-3.5 and Llama, on the knowledge probing task with the Bulgarian subset of the mLAMA dataset. We also plan to make use of the available answers in the EXAMS dataset for fact-checking evaluation. Additionally, we would like to further research if we can use other datasets from the bgGLUE benchmark (Hardalov et al., 2023) for our task.

3. Research and employ methods for prompt engineering that improve the multilingual performance of generative models, such as *cross-lingual-thought prompting*, proposed by Huang et al. (2023).

4. Finally, we would like to make our work more comprehensive by providing examples for different kinds of hallucination that occur in LLM-generated responses in Bulgarian.

Acknowledgments

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0008-C01

We would like also to thank our supervisor Dr. Trifon Trifonov for his kind support and suggestions regarding the paper topic.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Model	Response
text-davinci-003	Логиката изучава животните и начините по които се придвижват и решават проблеми. Тя извлича дразнителни и аналитични науки, за да помогне хората да се интелектуално подобрят.Тя може да бъде приложена в различни дисциплини и сфери, като право, економика , наука, философия и много
gpt-3.5-turbo-0613	Логиката изучава законите на рационалното мислене и следва стриктни правила, които уреждат процеса на извод и аргументация. Тя се занимава с анализа на размислите и тяхната структура, разкривайки начините, по които принципите на правилното резониране водят до верни изводи от зададени предпоставки.

Table 4: Example of responses different models give when prompted “Напиши абзац, започващ с ’Логиката изучава’:" *(“Write a paragraph beginning with ’Logic studies:’”)*. Non-factual and nonsensical statements are marked in red. Misused words and phrases, specific for the Bulgarian language, are marked in blue. Translations are provided in Table 7.

Example	Correct form	Automatic translation	Literal translation	Error type
економика	икономика	economics	-	spelling
да се интелектуално подобрят	да се развият интелектуално	to improve intellectually	-	word order and terminology
анализ на размислите	анализ на съжденията	analysis of thoughts	analysis of reflections	terminology
правилно резониране	правилно разсъждение	correct reasoning	correct resonance	terminology

Table 5: Examples of *”foreign language hallucinations”*.Automatic translations were done via Goodle Translate and DeepL translate

.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Melania Berbatova and Filip Ivanov. 2023. An improved bulgarian natural language processing pipeline.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating crosslingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

A. H. Miller P. Lewis A. Bakhtin Y. Wu F. Petroni, T. Rocktäschel and S. Riedel. 2019. Language models as knowledge bases? In *In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019*.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Ves Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgglue: A bulgarian general language understanding evaluation benchmark. *arXiv preprint arXiv:2306.02349*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’20, pages 5427–5444, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

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.

Shangqing Tu Shulin Cao Daniel Zhang-Li Xin Lv Hao Peng Zijun Yao Xiaohan Zhang Hanming Li Chunyang Li Zheyuan Zhang Yushi Bai Yantao Liu Amy Xin Nianyi Lin Kaifeng Yun Linlu Gong Jianhui Chen Zhili Wu Yunjia Qi Weikai Li Yong Guan Kaisheng Zeng Ji Qi Hailong Jin Jinxin Liu Yu Gu Yuan Yao Ning Ding Lei Hou Zhiyuan Liu Bin Xu Jie Tang Juanzi Li Jifan Yu, Xiaozhi Wang. 2023. Kola: Carefully benchmarking world knowledge of large language models.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *to appear in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Freedom Preetham. 2023. Mathematically evaluating hallucinations in llms like gpt4.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

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

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Appendix A Additional experiments

We aimed to run our experiments also with the large language models LLama (Touvron et al., 2023a) and LLama 2 (Touvron et al., 2023b), developed by Facebook. The company has not yet made a public application programming interface (API) available for these models, leading us to employ their minimal open-source software (OSS) version, LLaMa-7B, and its successor LLaMa-2-7B, on our local system. Our attempts to replicate the experiments encountered notable time constraints arising from hardware limitations, as the computations were performed on our local machine. Therefore, we decided to leave these experiments for our future work.

Appendix B Translations of Examples

Tranlsations of examples of model-generated text in Bulgarian are given in Table 6 and Table 7.

Extrinsic Hallucination	**Nonsensical Statement**
The capital of Bulgaria is the largest city in Europe, and over 1.5 million people live in it.	Bulgarians are the poorest in Europe, but they are the poorest in the world.

Table 6: English translations of the examples shown in Table 1.

Model	**Response**
text-davinci-003	Logic studies animals and how they move and solve problems. It derives provocative and analytical sciences to help people improve intellectually. It can be applied in various disciplines and fields, such as law, economics, science, philosophy and many
gpt-3.5-turbo-0613	Logic studies the laws of rational thinking and follows strict rules that govern the process of inference and argumentation. It deals with the analysis of thoughts and their structure, revealing the ways in which the principles of correct resonance lead to correct conclusions from given premises.

Table 7: Translated examples of Table 4.