Evaluating Transformers on the Ethical Question of Euthanasia

Gerold Schneider and **Giovanni Spitale**
University of Zurich
gschneid@ifi.uzh.ch, giovanni.spitale@ibme.uzh.ch

Abstract

We evaluate the ability of transformers to test if they can reliably predict the recom–mendation by a euthanasia commission on whether, given the report of the commission, euthanasia was justified. As evaluation strategy we prompt the transformer to give binary *yes/no* answer in chatbot mode, which we then compare to the final verdict on the 72 publicly available cases as gold standard.

Starting from a bag-of-words document classification baseline with logistic regression, which reaches 93% accuracy, we step up to a zero-shot BART MLNI model, Llama 2, OpenChat, Llama 3, and finally GPT-4.

The BART zero-shot model performs slightly below the baseline, Llama 2 fails to follow the prompts in all cases, while OpenChat beats the baseline. The results obtained on Llama 3 support the claim that it is possibly the currently best open model, although a single task with few instances provides limited evidence. GPT-4 does not make a single mistake, underlining its superior semantic detail, though with the same caveats.

1 Introduction

1.1 Large Language Models

Large Language Models (LLMs), particularly GPT-3.5 and GPT-4, which are the base of ChatGPT, are revolutionizing AI, content analysis and computational linguistics. They reach human levels of performance on many tasks (Strachan et al. 2024, Ronan and Schneider 2024). Also in medical applications, Beaulieu-Jones et al. (2023) attest near-human surgery knowledge, while Liévin et al. (2023) report that ChatGPT-3.5 has “human level performance” on answering multiple-choice medical exam questions (USMLE and MedMCQA). At the same time, they also give absurd answers in situations of sparse data, instead of admitting that they cannot answer (Zhang 2023). In critical applications, hallucinations could have disastrous consequences. We thus investigate texts from the highly ethical question of euthanasia, simulating a situation in which a chatbot takes decisions over death and life.

1.2 Risk of hallucinations

The risks posed by hallucinations are enormous in the situation that we simulate, due to the highly ethical nature of the question addressed in the data (Buiting et al. 2009). Decisions on euthanasia are literally questions of death and life, and blindly trusting AI could have catastrophic consequences. The performance on the baseline document classification and on zero-shot BART models (see Section 4) shows that the risk would also be high in reality.

2 Data and Methods

2.1 Data

The data used are the publications of the Dutch Euthanasia Commission reports (Regional Euthanasia Review Committees 2017) and we use the subset of 72 reports available in English[1] that are made publicly available. The reports assess if a physician complied with the strict rules that are laid down by law, the so-called due care criteria. Physicians who fail to observe these statutory requirements could be criminally liable.

[1] https://english.euthanasiecommissie.nl

An example of a report is given in the following passage, example (1).

(1) *The patient, a man in his eighties, had suffered for 10 years from macular degeneration (which causes cells in the centre of the retina to die) in both eyes, which caused his eyesight to deteriorate. Around the same time, an obstructed blood vessel in the retina caused blindness in his right eye. Six months before his death, his left eye deteriorated so much, despite the start of treatment, that he was no longer able to read, even using aids. In addition to these eyesight problems, he was uncertain when walking, which was aggravated by his near-blindness. In recent years he had become unwell and fallen several times. Because he had become almost totally blind, the patient could no longer read (which was extremely important to him) or pursue his other hobbies. He was suffering from the loss of these activities, which were essential to him. He also suffered from the loss of self-reliance caused by his impaired vision, and the fact that he knew that there was no prospect of improvement whatsoever. The patient, who had always had a wide range of interests and a great intellectual appetite, experienced his suffering as unbearable. The committee found that the physician had plausibly argued that he was reasonably able to conclude that the patient's suffering was unbearable to him and without prospect of improvement, and that it was unlikely that optical aids and possibly surgery would enable him to read again. The other due care criteria were also fulfilled.* (https://english.euthanasiecommissie.nl)

2.2 Methods

Our method employs a range of Language Models: we first use a bag-of-words document classification baseline with logistic regression, a classical supervised machine learning scenario. Then, we test several LLMs, starting with BART MNLI, then Llama 2, OpenChat, Llama 3, and finally GPT-4. We prompt the LLMs in chatbot mode to provide *yes/no* answers. On the one hand, this prevents verbose hallucinations, on the other hand it allows us to evaluate LLMs with precision, recall, and F-scores, like in a supervised scenario. We rely on the latest LLMs and show in our evaluation how the risk reduces as a function of the complexity of the LLM. These results are given in section 3.

In section 4 we describe our error analysis stage. We query the chatbot to provide arguments for its decision, which allows one to zone in on the source of error, thus offering partly explainable AI despite the intrinsic blackbox character of LLMs.

2.2.1 Document Classification

As baseline, we use logistic regression with L2 regularization and 10-fold cross-validation on the 72 euthanasia texts, in a classical supervised binary document classification scenario, by means of the tool *LightSide*[2]. Because document classification is a linear method, we can interpret the features with the strongest weight as particularly good discriminators.

2.2.2 BART MNLI

Our approach employs the BART multilingual natural language inference LLM provided by Facebook[3]. We test in a zero-shot scenario if the inference to the commission's verdict of yes or no is more likely. Then, we compare the probabilities of the inference "The committee agrees that euthanasia was justified" for *yes*, versus "The committee thinks that euthanasia was not justified" for the *no* class. We tested 5 further prompt variants, but obtained worse or equal results.

Observe that in a zero-shot setting, the method cannot learn from the provided data, and has to rely completely on its general world knowledge. In order to test the general knowledge of the Facebook BART MLNI model on euthanasia, we asked it definitional questions; the dialogue is given in Figure 1.

The assessment shows that BART has a reasonable general knowledge on euthanasia, though inevitable cultural biases and stance in the training data emerge. With a probability of 3.58%, euthanasia is defined as *murder*. The use of web training data from different cultural backgrounds may well have caused higher or lower probabilities.

---

[2] https://github.com/LightSideWorkbench/LightSide

[3] https://huggingface.co/facebook/bart-large-mnli

```
input_text="euthanasia is";
candidate_labels = ["assisted suicide","death","murder","the end of life",
                    "an illness","a patient","a famous painter","a Greek godess",
                    "an ethical dilemma","an ethical question","a political debate",
                    "connected to suffering","a consequence of suffering",
                    "a consequence only permitted for extreme suffering","a reason for suffering"]
```

```
# multi_label=True will return confidence score for both labels independently
model_dict = classifier(input_text, candidate_labels, multi_label=True)
# Zip results to dict
result_dict = dict(zip(model_dict.get('labels'), model_dict.get('scores')))
```

```
result_dict
```

```
{'assisted suicide': 0.9878346920013428,
 'the end of life': 0.9851932525634766,
 'connected to suffering': 0.9771237373352051,
 'death': 0.9526866674423218,
 'an ethical dilemma': 0.9075977802276611,
 'a consequence of suffering': 0.8390344977378845,
 'an ethical question': 0.8190814256668091,
 'a consequence only permitted for extreme suffering': 0.6314218044281006,
 'a reason for suffering': 0.3726317286491394,
 'a patient': 0.11474824696779251,
 'a political debate': 0.06330907344818115,
 'murder': 0.03578542172908783,
 'an illness': 0.033096276223659515,
 'a Greek godess': 0.0007784005138091743,
 'a famous painter': 0.00019802608585450798}
```

Figure 1. Answers of BART MLNI to definitional questions on euthanasia, sorted by decreasing probability.

2.2.3 Llama, OpenChat and GPT-4 models

We used chatbot prompting for the remaining transformer models, also in a zero-shot setting. For this, we relied on *ollama* to run the Llama (Touvron et al. 2023) and OpenChat (Wang et al. 2024) models on a local computer[4]. As for ChatGPT, the Open AI API with GPT-4 (OpenAI 2023) was used.

In the following dialogue (2), after assigning a given report to the variable X, we first prompt ChatGPT (and equally the other models) to give a *yes*/*no*-answer, then we ask for an explanation. The report X has been replaced by 3 dots for brevity, example (1) shows one (different) instance.

(2) ***User****: The background text X from a discussion on Euthanasia is "…". Given X, does the committee agree that euthanasia was justified? Please answer just with yes or no.*
ChatGPT*: No.*
User*: Why?*
ChatGPT*: In case X, the committee did not agree that euthanasia was justified because the physician did not adhere to the established euthanasia procedures and guidelines. Specifically, the physician administered a lower dosage of the coma-inducing substance than recommended and did not adequately check the depth of the patient's coma before administering the muscle relaxant. This could have resulted in the patient being in an insufficiently deep coma, meaning they might have been able to perceive the effects of the muscle relaxant. Therefore, the committee concluded that the physician did not exercise due medical care when terminating the patient's liefe, which is a requirement for euthanasia to be considered justified and legally compliant.*

3 Results

In order to mitigate and assess hallucinations, we force LLMs to provide *yes/no* answer. This allows one to evaluate their performance in comparison to the gold standard of the final recommendation of the Euthanasia commission.

Yes/no prompting renders evaluation as easy as in supervised learning whenever a gold standard is available. The results are given in Table 1.

Table 1. Performance of the tested language models on prediction of Euthanasia recommendations

Model	Correct	Incorrect	Accuracy
Log.Reg.	67	5	93%
BART MLNI	63	9	88%
Llama2	NA / 67	NA / 5	NA/93%
OpenChat	68	4	94%
Llama3	71	1	99%
GPT-4	72	0	100%

While BART MLNI performs slightly below the document classification baseline, all other LLMs perform above the baseline. In particular, we observe that GPT-4 does not make a single prediction mistake, and that Llama 3 is considerably better than Llama 2. Llama 2 refuses to observe the prompt and does not want to commit to a *yes*/*no* answer in two contested cases. If we count both cases as incorrect, Llama 2 obtains an accuracy of 93%. We need to point out though, that our evaluation set of only 72 texts is too small to provide trustworthy answers. While the fact that GPT-4 does not make a single error is impressive, this may partly be an effect of the small size of the evaluation data.

[4] https://www.ollama.com

4 Discussion

As mentioned above, the data size is small. It can also be argued that the task is not very difficult. In most cases, it is clear to humans what the recommendation will be, since the commission's report was written as an explanation of the recommendation. The recommendation is written after the euthanasia, and as the expert commission has no legal power, the recommendations are thus not verdicts in any legal sense.

Nevertheless, our experiment has shown that AI has reached the level of human understanding of demanding argumentative texts. Further, the methods are easy to use and – except for ChatGPT with GPT-4 – easy to install (Section 4.1), point out ethical dangers (Section 4.3), and we can use AI itself to curb hallucinations by prompting and by querying (Section 4.2).

4.1 Prerequisites for the approach

Document classification with supervised logistic regression bag-of-words models need minimal prerequisites and runs extremely fast. Next, BART MLNI runs efficiently in all recent Python environments. The open source chatbot models were run via *ollama* which is very easy to install. The 7 billion parameter models that we used have the important advantage that they can be run on recent powerful desktop or laptop computers, we used a MacBookPro with 32 GB RAM and an M2 processor. Reaction times were 1 to 5 seconds per query, which means that for tasks involving up to a few thousand instances, they are easily applicable. Figure 2 shows an R session using Llama 3.

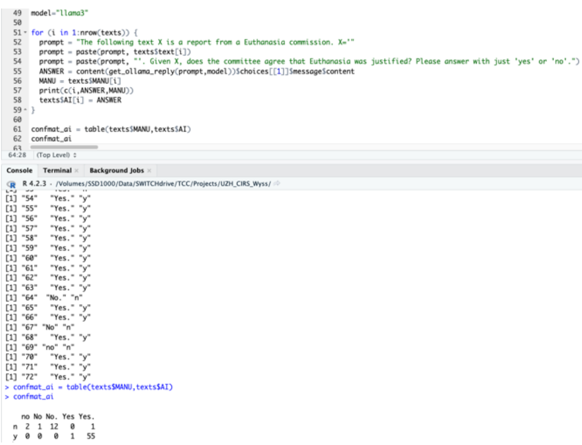

Figure 2. R Session querying the *ollama* API on the euthanasia dataset

By contrast, GPT-4 cannot be run locally, the per-token costs grow forbidding when used on thousands of cases. Reaction times are similar to those in the *ollama* models. Each of the tested approaches need minimal programming skills in Python or in R.

4.2 Explainable AI?

ChatGPT allows users to query for detailed reasons that lead to the classification suggested by the generative LLM, as illustrated in example (2). They thus potentially increase explainability in an intrinsic blackbox situation.

However, there are unexpected dependencies. First, very small changes in the parameter space lead to opposite results. In the one case of euthanasia reports, in which Llama 3 gave an incorrect prediction of the outcome, we wanted to assess the reason for the erroneous prediction. We thus extend the prompt by one line, "Then, give a brief explanation of your answer". This tilted the answer from the incorrect prediction of *yes* to the correct *no*.

Second, the chatbot models (Llama, OpenChat, GPT-4) are not able to deliver reliable confidence scores. To extend the above prompt by "Then, tell me how sure you are about your answer in terms of a probability between 0 and 1." instead of the brief explanation equally triggers the correct answer *no*, with an alleged confidence of 99%. A screenshot of the example session with the contested case 53 is given in Figure 3.

In this respect, the confidence scores of BART, as given in Figure 1 for the definitional questions of euthanasia, are more reliable, and so are the weights offered by linear models, for instance the word and *n*-gram weights given by the baseline document classification model.

Raising the temperature to above zero also sometimes leads to opposing results, for example in the contested case 53. Comparing distributions of these results may lead to a more reliable assessment of model confidence.

4.3 Challenges in applying this method

Given the highly ethical nature of the question of Euthanasia, trusting the decisions of LLMs would be extremely dangerous. Also using them as advisors raises questions on possible biases. They range from the question of how similar *euthanasia* and *murder* are according to our training data

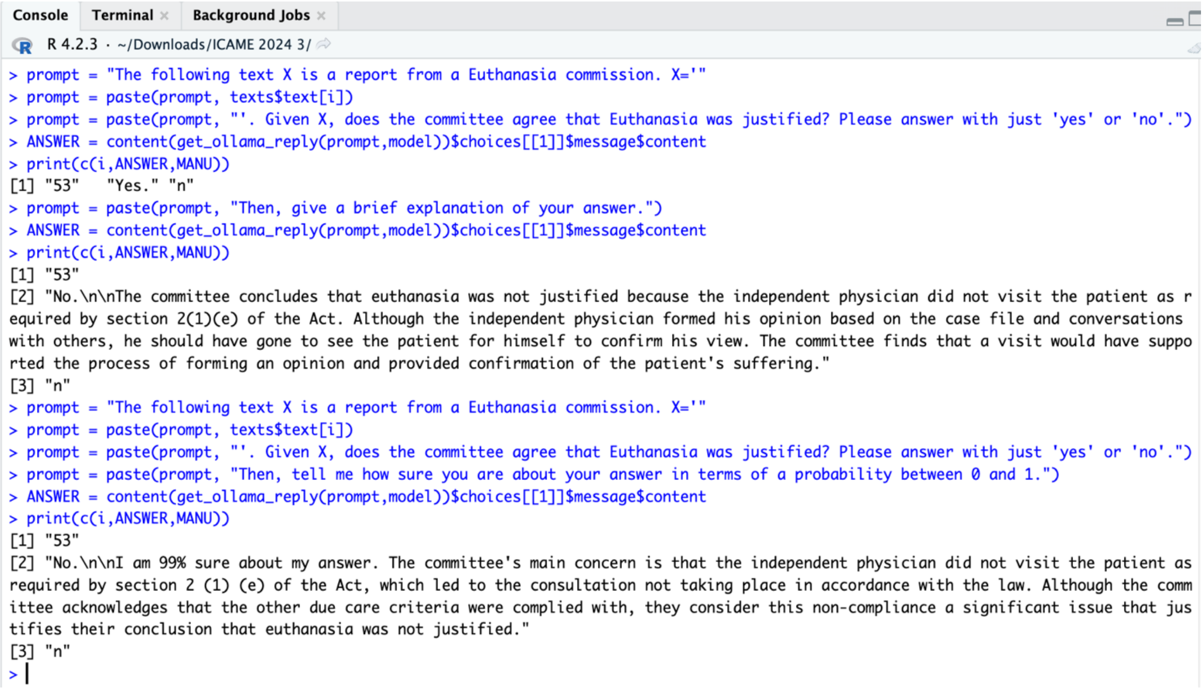

Figure 3. Asking for explanations or for a confidence score tilts the answer in Llama 3. The delivered very high confidence score is not trustworthy, given that omitting a call for explanation leads to the opposite answer

(BART MNLI says 3.58%) to the fact that trivial changes in the prompt tilt the answer of Llama 3.

But we hope to have contributed to showing on the one hand that human or near-human performance on deep-semantic tasks is now possible, while on the other hand LLMs make mistakes, and error analysis is difficult, as the models partly fail to deliver confidence scores. Constant, careful evaluation remains pivotal.

5 Conclusions

We have illustrated that prompting LLMs to provide *yes/no* answer allows one to evaluate their performance in comparison to the gold standard of the final recommendation of the Euthanasia commission. *Yes/no* prompting renders evaluation as easy as in supervised learning whenever a gold standard is available. Additionally, Chatbot settings allow users to query for detailed reasons that lead to the classification which is suggested by the generative LLM. This approach thus increases explainability in an intrinsic blackbox situation.

Comparing the different models, the BART zero-shot model performs slightly below the baseline, Llama 2 fails to follow the prompts in all cases, while OpenChat beats the baseline. The results obtained on Llama 3 support the claim that it possibly is the currently best open model, although a single task with few instances provides limited evidence. GPT-4 does not make a single mistake, underlining its superior semantic detail.

However, a number of caveats are needed. Very subtle changes in the prompt can lead to different answers, even if temperature is set to zero. In one case, adding a sentence asking for an explanation for a misclassification tilted the answer itself. In this sense, LLMs remain blackbox models, and evaluation is pivotal.

Acknowledgements

We are grateful to the projects Digital Religion(s)[5] and Mind the Patient Lab[6] at University of Zurich, which partly supported this research.

References

Hilde Buiting, Johannes van Delden, Bregje Onwuteaka-Philpsen, Judith Rietjens, Mette Rurup, Donald van Tol, Joseph Gevers, Paul van der Maas

[5] https://www.digitalreligions.uzh.ch/

[6] https://www.dsi.uzh.ch/en/research/projects/dsi-lab-infra/mind-patient-lab.html

and Agnes van der Heide. 2009. Reporting of euthanasia and physician-assisted suicide in the Netherlands: descriptive study. *BMC Medical Ethics* 10(18). https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-10-18

Brendin R. Beaulieu-Jones, Sahaj Shah, Margaret T. Berrigan, Jayson S. Marwaha, Shuo-Lun Lai and Gabriel A. Brat. 2023. Evaluating capabilities of large language models: performance of GPT4 on surgical knowledge assessments. *medRxiv*. doi: 10.1101/2023.07.16.23292743

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt and Ole Winther. 2023. Can large language models reason about medical questions? *Patterns* 5(3), 100943. doi: 10.1016/j.patter.2024.100943.

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

Regional Euthanasia Review Committees. 2017. *Judgments - Regional Euthanasia Review Committees*. Available online at: https://english.euthanasiecommissie.nl/judgments (accessed June 30, 2024)

Patricia Ronan and Gerold Schneider. 2023. Can ChatGPT solve a Linguistics Exam? *arXiv* 2311.02499.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*. doi: 10.1038/s41562-024-01882-z

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, Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, Yang Liu. 2024. Openchat: Advancing Open-Source Language Models With Mixed-Quality Data. *arXiv* 2309.11235v2

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.