Comparison available open source LLMs

**Introduction**

There are civil servants who use language models like Chat-GPT to improve their work, but there are also concerns about the fact that the government has no control over these models and becomes dependent on large companies. That's why some people wonder if we can create our own model within the government, one that keeps the data internal and is specifically trained for the tasks we want to use it for. This possibility is what we want to explore in this project.

Starting from scratch is unrealistic; it takes an enormous amount of time, money, and computational power to build a model, and we lack the expertise for it. However, what we can attempt is to download an open-source model and adapt it for internal use. Open-source models are freely available for download, and their code and functionality can be adjusted locally. These changes and the data used for them remain on-premises.

It's important to realize that the available models and techniques are rapidly evolving, and the debate surrounding them is continuously developing. This document describes recent developments in the field of open-source models as of September 2023. The rapid progress means that by the time a decision is made to use open-source models within the government, new models are likely to be better suited by that time. It's even conceivable that during the course of this project, new developments may require us to adjust our approach.

**Foundational models: LLama2.**

As mentioned, training a model from scratch is costly. Only a handful of companies and organizations have undertaken the basic training of a model and made their model open source. Such a model is referred to as a foundational model. Not every open-source model is a foundational model. Some people download a foundational model, further train it themselves, and then make the result publicly available. Even if you want to use such a model, it's important to know which foundational model it was based on, as the data and methodology used to train the foundational model always have an effect on the ultimate performance.

MPT (trained by the company Mosaic) and Falcon models (trained by the Technology Innovation Institute (TII), a research center funded by the Abu Dhabi government) are examples of open-source foundational models. The Falcon models are trained in English, German, Spanish, and French, while the MPT models are trained only in English and are relatively proficient in programming.

Since July 19, 2023, new open source models made by Meta are available. These models, called LLama 2, generally seem to outperform the MPT and Falcon models. Furthermore, unlike Mosaic and TII, Meta has also released a paper detailing exactly how it was trained. This paper explains that the LLama 2 models were not only trained on a wealth of publicly available data but also underwent extensive "reinforcement learning with human feedback" (RLHF) and "red-teaming." RLHF involves people assessing the model's responses, based on the helpfulness and safety of the answers. Safety, in this context, means that the model does not provide harmful responses, such as responses containing incorrect information, harmful advice or responses that assist users in engaging in illegal activities. Based on the numerous assessments by the individuals participating in RLHF, the model is further trained to provide helpful and safe responses. In red teaming, people actively try to get the model to perform unsafe actions and use clever techniques to push its limits. This tests the model's robustness, and every time the model makes a mistake, its weights are adjusted to reduce the chances of similar errors in the future. However, a significant drawback is that LLama 2 is trained only in English, which makes both its performance and the effects of safety training in Dutch uncertain.

There exists an LLama 2 model with 7, 13, 34, and 70 billion parameters. In general, the more parameters a model has, the more computational power (and therefore more money and energy) you need to download, adapt, and use the model. At the same time, more parameters often means better performance. Given that this is a research project, in which the aim is to test what is possible at a small scale, we strongly prefer models with fewer parameters. Furthermore, the government generally aims to be energy-efficient and fiscally responsible, so should only consider more parameters if it significantly improves the results.

Therefore, the Falcon model that was released shortly after Llama 2 is not of interest to us. ITT has released a model that, according to some metrics, performs slightly better than the 70 billion-parameter LLama 2 model but has 180 billion parameters, which is significantly more.

The French company MistralAI released a model with only 7 billion parameters on September 27, 2023, which performs better in many tests than the 13 billion-parameter LLama 2 model. This is quite interesting for our purposes. The exact mix of languages used in its training is unknown, but based on our personal experiences it seems to perform slightly better in Dutch as well. Moreover, MistralAI states that they will soon release models that can handle multiple languages. If this happens in time to include it in this project, it also has potential. This demonstrates how quickly this field is evolving; when we initially drafted this document, this model did not exist, and now it is our top preference.

**Table: foundational models compared**

|  | Llama 2 | Falcon | MPT | Bloom | Mistral 7B |
| --- | --- | --- | --- | --- | --- |
| Trained by | Meta (facebook) | Technology Innovation Institute (financed by the government of Abu Dhabi) | Mosaic | Big science: Collaboration of >1000 AI researchers | Mistral\_AI (6 month old paris based startup) |
| Number of parameters in which it is available | 7 billion / 13 billion / 70 billion | 7 billion / 40 billion / 180 billion | 7 billion / 30 billion | 7 billion / 176 billion | 7 billion |
| LLM leaderboard score | 54.32 / 58.66 / 67.35 | 47.01 / 61.48 / 68.74 | 47.38 / 56.15 | 42.07 / niet opgenomen | 62.4 |
| Languages it has been trained and tested in | English | English, German, Spanish, French | English | 46 languages, no Dutch | Unknown |
| Safety training | Extensive RLHF and red teaming | Unknown | Unknown | Unknown | Unknown |
| Paper | [[2307.09288] Llama 2: Open Foundation and Fine-Tuned Chat Models (arxiv.org)](https://arxiv.org/abs/2307.09288) | Only regarding the dataset: [[2306.01116] The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only (arxiv.org)](https://arxiv.org/abs/2306.01116) | No paper, just a blog: [MPT-30B: Raising the bar for open-source foundation models (mosaicml.com)](https://www.mosaicml.com/blog/mpt-30b) | [[2211.05100] BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (arxiv.org)](https://arxiv.org/abs/2211.05100) | No paper, just a blog:  [Mistral 7B | Mistral AI | Open source models](https://mistral.ai/news/announcing-mistral-7b/) |

The LLM leaderboard score is obtained from the Hugging Face platform, which is a platform where people make their models open source. It represents a weighted average of how well models perform on commonly used tests that assess various aspects such as knowledge, reasoning abilities, programming skills, and common-sense logic. These scores may not be a perfect reflection of how well the model will perform in the specific tasks we intend to use them for, but they do provide some insight into the overall competence level of a model.

**Getting Trained by GPT-4: Orca.**

Microsoft has recently come up with a new technique to enable models with relatively few parameters to achieve results that come close to models with an extremely large number of parameters. The idea is that you can improve a small open-source model by learning from the best and largest models available, such as GPT4.

They have created a dataset called Orca, which consists of inputs combined with the outputs that GPT-4 would provide. This dataset can be used to teach a model how GPT-4 would respond to a prompt. This dataset is freely available online. In a paper Microsoft describes how they trained LLama-2 with the Orca dataset. They showed that you can get a model to perform much better than expected based on the number of parameters by using this approach. With a fraction of the computational power that GPT-4 required to achieve its competence, a significant amount of knowledge and capability can be transferred. The Orca dataset also demonstrates that less data is often needed to train a relatively competent model than previously thought if the data is smartly constructed.

Additionally, some individuals have trained the 7 billion-parameter Mistral model with the Orca dataset, effectively having Mistral trained by GPT-4. The result is Mistral-7B-OpenOrca, a model that seems to perform very well for its size of 7 billion parameters. This model is the one that will serve as the basis for our training. There is a demo available where you can test the model yourself. At first glance, it seems to work better than the alternatives, in this sense our personal experience aligns with the official test results.

**Training in Dutch: Conversation with Bram Vanroy.**

Bram Vanroy is a postdoctoral researcher with expertise in natural language processing (NLP) and machine translation at the Catholic University of Leuven (KU Leuven). He is also an active contributor to open-source projects and has developed datasets, models, and demonstrations related to NLP, particularly in Dutch.

Bram has trained foundational models multiple times with Dutch-language data and has made the resulting models open source and publicly available on Hugging Face. He has, for example, trained the 13 billion-parameter LLama-2 model and the 40 billion-parameter Falcon model in Dutch. He mentioned that the results for LLama-2 were not optimal, and it is challenging in practice to teach a model that had very little pre-training in Dutch to perform well with the currently available datasets. Falcon 40 billion performed somewhat better, but it also has more parameters. He advised us that it would be better to make use of the Mistral models than his own.

He also noted that the currently available Dutch datasets are of low quality. His approach had always been to use as much data as possible. However, the Orca project has shown that you can achieve excellent results with relatively little data, if the data is of high-quality. Vanroy suggested that it would likely be of great value for the development of open-source language models in Dutch to translate the entire Orca dataset into Dutch. However, he did not have the budget for this himself (he estimated about 5000 euros in computational power, plus expertise for translation and using the translation for training). If there is a widespread desire within the Dutch government to have language models that are proficient in Dutch, it may be worth translating the Orca dataset and making it publicly available. The question then arises as to who would be best suited to carry out this task.

**Conclusion.**

The field is rapidly evolving, so this document will quickly become outdated. As of the time of writing, October 5, 2023, Mistral-7B-OpenOrca seems to be the best foundational model to further train for chatbot functions within the Dutch government. This is mainly because the performance-to-parameter ratio is very good. This is important to us because more parameters mean more computational power, and thus more cost and energy consumption. The Mistral 7 billion-parameter model is the best foundational model with 13 billion parameters or less. Furthermore, the Orca dataset helps models with far fewer parameters than GPT-4's hundreds of billions to achieve a significant portion of its performance.

Currently, there are no models available that are explicitly trained in Dutch with similar quality. A translation of the Orca dataset could be beneficial but would also require funding and expertise.