Domain shift in distilled T5-based models for language translation tasks

Giovanni Novati

Computer Science Master's Course, Universitá degli Studi di Milano.

Abstract

The abstract serves both as a general introduction to the topic and as a brief, non-technical summary of the main results and their implications. Authors are advised to check the author instructions for the journal they are submitting to for word limits and if structural elements like subheadings, citations, or equations are permitted.

Keywords: distillation, domain shift, language translation, t5, mt5

1 Introduction

In recent years, more and more cumbersome LLMs have come to life in the AI race led by OpenAI and joined by all other competitors and big techs like Google, Anthropic, and Meta, just to say some. These companies have created models with billions if not trillions of parameters [1], and are general purpose, which means they are capable of doing a wide range or tasks such as natural language understanding and generation, decision making and reasoning [2].

The main downside of these models is also their main advantage: the size. Running inference requires expensive hardware in order to load all parameters into dedicated GPU memory. Moreover, if we want to perform a task that only uses a subset of the model capabilities, all the parameters are still active, leading to a waste of memory and computational power.

In the case we only want to perform a certain task, it might be useful to condense a larger model into a smaller one, specialized on that specific task while preserving comparable performance. This process is commonly known ad distillation [3], and consists in transferring knowledge from a teacher model into a smaller student model.

A limitation arises when the teacher model we want to use is proprietary, and internal parameters or output logits are not accessible. Using a black-box makes the distillation process more difficult, and in those scenarios alternatives strategies must be considered. One of them is training the student using only teacher's predictions; in this way we are still able to train the smaller model, but a reduction in performance is expected compared to full logits distillation [4].

In this paper, I investigate the distillation of task-specific student models from a GPT-5 Mini teacher using only its output predictions. I test two different student architectures across multiple parameter counts in order to assess their performance under domain shift. The models are trained on English-to-Italian translation and evaluated using BLEU and chrF++ metrics to measure both in-domain and out-of-domain effectiveness.

2 Methodology

2.1 Task

First, let's define the task of interest. I wanted to train the model on something useful and challenging, in order to stress the student and understand its limitations. I opted for a translation task, specifically English to Italian, since translating into morphologically richer languages is a more difficult task [5] compared to the opposite. Italian is a more complex language as it has more verb tenses and information like gender and number is embedded into words. These elements are not always present into English sentences and must be inferred based on the context.

2.2 Datasets

After that, I needed to define two domains used to train and evaluate the students. Dataset A [6] contains 175.622 couples of English sentences with their related French translation. These sentences cover everyday life situations, with an average length of 42 characters. The second one, Dataset B [7], contains 1.534.699 english sentences with an average length of 133 characters, and are related to educational and academic contexts. Note that the author of the latter one says that the dataset may introduce some biases, like a more formal writing style. Table 1 shows examples from both datasets.

I then extracted about 35.000 sentences for training and 5000 for testing from each dataset [8], and created the final datasets used for the experiments. Since Dataset A

Source	Example sentences
Dataset A	I'll get something to drink for both of you. I have to do my best. I'm going to wear these shoes on our date tonight.
Dataset B	Muscles, tendons, and ligaments depend upon proper joint movement to function at optimal levels. It is advisable to keep water levels some distance below where the tiles are to prevent any damage. Technological development has also contributed to long-term efficiency and productivity.

Table 1 Comparison of sample sentences from Dataset A and Dataset B.

consists of English-French pairs, I discarded the French side and kept only the English sentences.

Next, using OpenAI API, I prompted GPT-5 Mini to translate all the english sentences into Italian, and saved the resulting source-target pairs in a CSV file. After completing the translations, I extracted an equal number of samples from Dataset A and Dataset B and merged the two into a new Dataset AB, which combined the two domains. I chose GPT-5 Mini in order to keep costs low; using a more advanced model such as GPT-5.2 would have increased the total cost by approximately 10 times.

2.3 Training

3 Results

4 Conclusions

References

[1] Abacha, A.B., Yim, W.-w., Fu, Y., Sun, Z., Yetisgen-Yildiz, M., Xia, F., Lin, T.: Medec: A benchmark for medical error detection and correction in clinical notes. In: Findings of the Association for Computational Linguistics: ACL 2025 (2025)

[2] Ye, Y., Zhang, Z., Ma, T., Wang, Z., Li, Y., Hou, S., Sun, W., Shi, K., Ma, Y., Song, W., Abbasi, A., Cheng, Y., Cleland-Huang, J., Corcelli, S., Goulding, R., Hu, M., Hua, T., Lalor, J., Liu, F., Luo, T., Maginn, E., Moniz, N., Rohr, J., Savoie, B., Slate, D., Webber, M., Wiest, O., Zhang, J., Chawla, N.V.: LLMs4All: A Review of Large Language Models Across Academic Disciplines (2025)

[3] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[4] Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., Yan, B., Chen, Y.: Survey on knowledge distillation for large language models: methods, evaluation, and application. ACM Transactions on Intelligent Systems and Technology (2025)

[5] Chahuneau, V., Schlinger, E., Smith, N.A., Dyer, C.: Translating into morphologically rich languages with synthetic phrases. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013)

[6] Kelly, C.: Language Translation (English-French). Kaggle (2020). https://www.kaggle.com/dsv/1067156

[7] Agentlans: High-Quality English Sentences. Hugging Face (2023). https://huggingface.co/datasets/agentlans/high-quality-english-sentences

[8] Vieira, I., Allred, W., Lankford, S., Castilho, S., Way, A.: How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In: Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track) (2024)