Distilling Knowledge for Assertion Generation: Alpha-Temperature Tuning in Smaller Language Models

Delft University of Technology

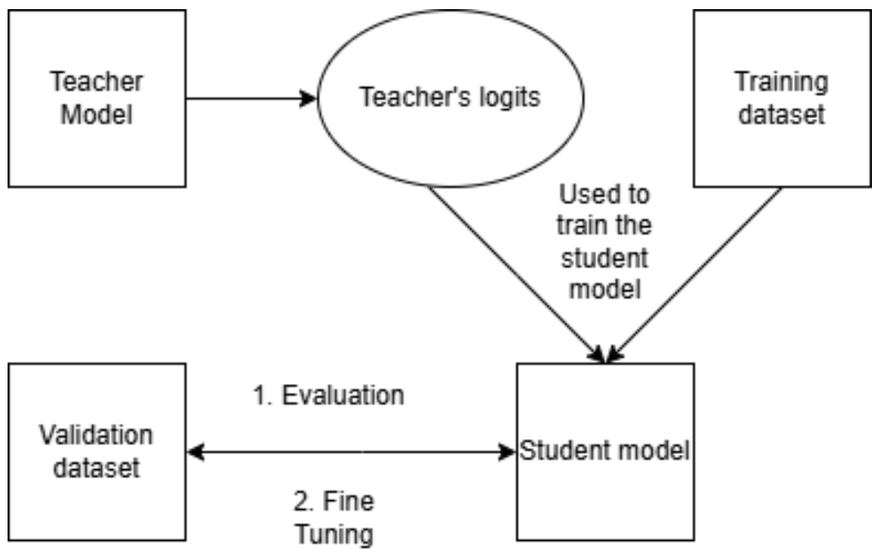
Author: Kristian Hristov
Email: K.G.Hristov@student.tudelft.nl
Supervisors: Mitchell Olsthoorn, Annibale Panichella

Introduction

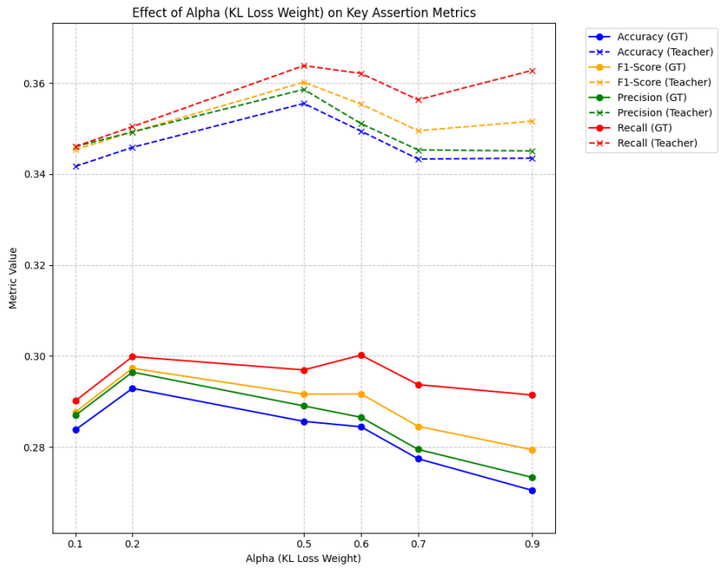
Software, and more specifically in our case – assertion-based testing, is vital because it ensures that our product is of high quality and is working as intended. However, it is very time-consuming and can delay software releases.
Large Language Models (LLMs), with their good understanding of code and natural language, emerge as a potential solution. But they come with their own limitations – high computational cost, latency issues, a dependance on a strong internet connection and inability to be used locally because of their large size.

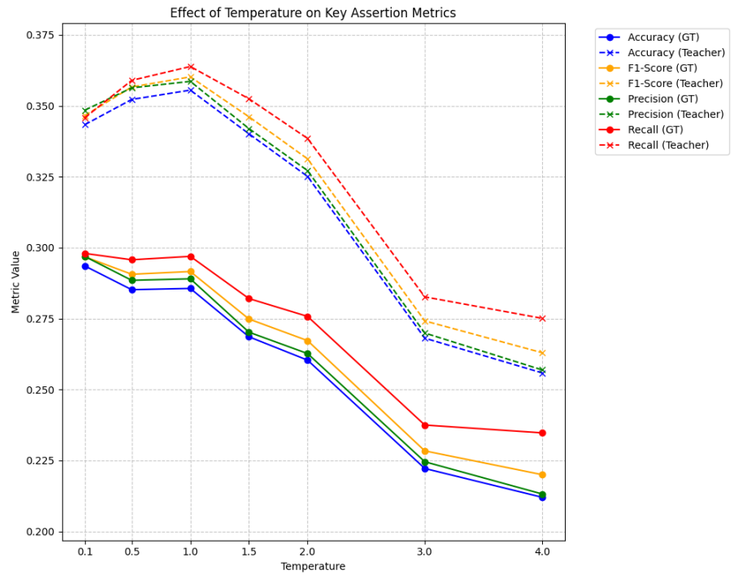
Knowledge distillation

A potential solution to our problem is knowledge distillation. Knowledge distillation is technique that offers an interesting approach to model compression. The core idea behind it is to use the knowledge of a large model, the teacher, to train a smaller one, called the student model. The student model learns both from the teacher one (soft target) as well as from a ground-truth data (hard target). The main trade-off is between the smaller size of the student model and its weaker performance when compared to the teacher one. We aim to investigate if this trade-off is worth and what are some optimal hyperparameters.



Results





As can be seen from the above graphs, the optimal values for our chosen hyperparameters depend on the goal we have. However, a potential optimal values for two of the hyperparameters – the alpha and the temperature are 0.5 and 1.0, respectively. We consider them promising values as they offer the best results when it comes to the student model performance when compared to the teacher one. The best student model we created is 4x smaller than its teacher and correctly reproduces ~36% of the teacher's assertions and ~30% of ground-truth assertions. Although the correctness of the student is below what people would want, the picture gets better when we consider that it is only 25% of its teacher's size. Our results call for further research on this topic.

Limitation and Future Work

Our study was constrained by the use of compressed teacher logits to manage data size. This inherently limited the information transferred from the teacher, capping the student's maximum potential performance.
Future work would ideally include: re-evaluating using original teacher logits, extend the evaluation to different programming languages, experiment with different student model architectures and focus of the balance of trade-off between size and performance.