ASPETTANDO LA
NOTTE EUROPEA DEI RICERCATORI

Optimization in Large Language Models: Innovative Approaches Leveraging Knowledge Distillation and Pruning

Paolo Italiani

Department of Computer Science and Engineering
PNRR PhD in "Computer Science and Engineering" - XXXVIII Cycle
Supervisor: Prof. Gianluca Moro

Background

Knowledge Distillation: In the NLP community, many works apply Knowledge Distillation (KD) to text classification tasks by training the student to mimic the teacher's output. On the other hand, other works use a frozen instruction-tuned LLM that generates artificial examples that are used as knowledge to train a student SLM with an order of magnitude fewer parameters. However, previous work on KD has focused mainly on text classification, neglecting generative tasks such as QA.

Token Pruning: Initial attempts to mitigate the computational complexity of transformers, achieved through the removal of non-informative tokens, were reliant on attention scores. PoWER-BERT uses a scoring function derived from attention scores to drop tokens based on their impact on others. Yang et al. (2022) adaptively determined the quantization precision levels of the tokens (i.e. 0 bit, 4 bit, and 8 bit) based on their importance, as gauged by their attention probabilities. Despite their popularity and promising outcomes, these solutions are grounded in predefined heuristics. Other approaches leverage Reinforcement Learning (RL) to train a policy network responsible for choosing the top-k tokens, therefore necessitating the formulation of a specific objective function.

Project Goals

Comprehensive knoledge distillation: Formulate KD solutions tailored for relatively unexplored domains, such as medical and legal, as well as tasks like question-answering, summarization, and named entity recognition.

End-to-end token pruning: Design token pruning techniques where the selection of the most crucial tokens is performed in an end-to-end fashion.

Experimental Approach

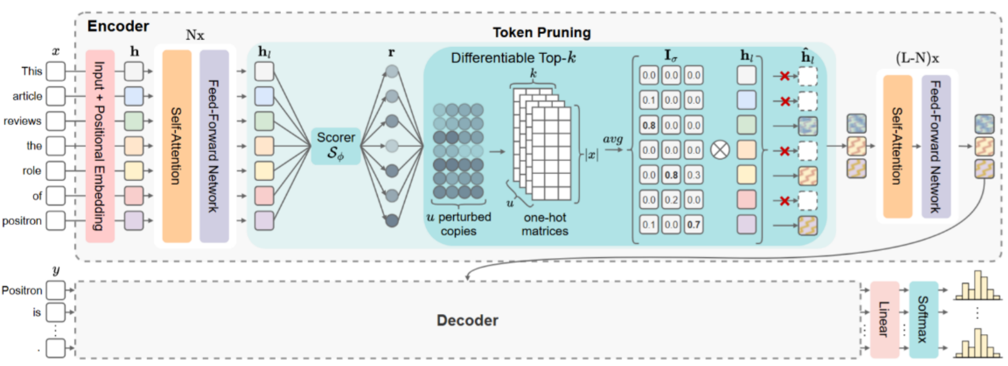
Ace-Attorney: a new legal framework designed to produce LQA-specific datasets and supervised SLMs through LLM KD. Precisely, given an input textual prompt, a state-of-the-art frozen LLM generates artificial samples that are used as knowledge to train a reduced-parameter student model. We collect LLM-crafted samples and create Syn-LeQA, our synthetic dataset proposed to address the scarcity of public LQA corpora. We also propose the Selective Generative Paradigm (SGP), which allows the generator and the retriever to work jointly to select the correct document to answer the question.

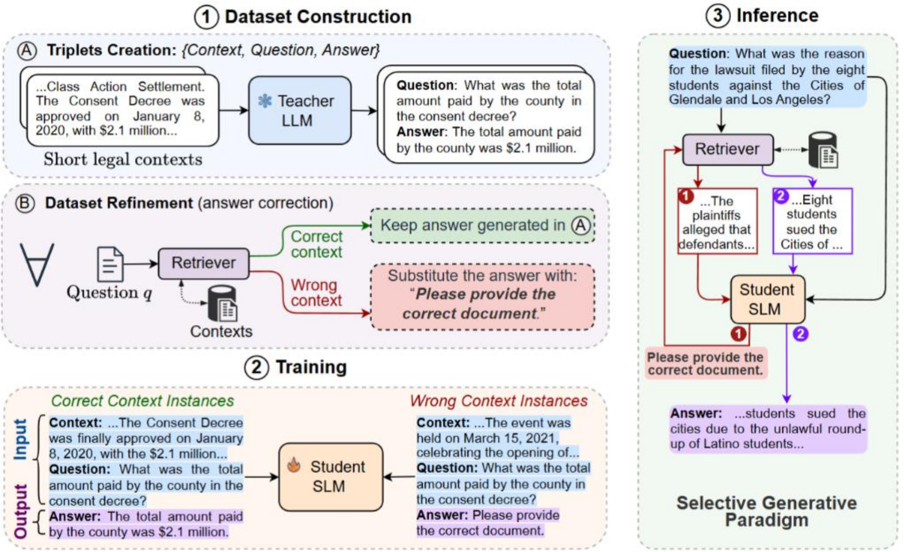
PrunePert: we use perturbated-topk, adeptly tackling the token pruning challenge through a differentiable top-k function. This approach enables the selection of the most crucial tokens in an end-to-end manner without the need for additional ad-hoc losses. Beyond the advantage of directly training the scorer on the final summarization loss, in contrast to prior works focused solely on efficiency during inference, our solution also optimizes resource utilization during training, as the top-k token selection occurs also at training time.

Expected Outcomes

Accessibility: Make advanced NLP tools usable on low-resource devices, improving access to legal, medical, and educational services.
Sustainability: Lower the environmental footprint of large-scale AI by reducing energy consumption.
Equity: Democratize AI applications in sensitive sectors, fostering fairer and more efficient access to knowledge.

Left: PrunePert architecture. Tokens are pruned by scoring their importance with a lightweight network and keeping only the top-k most relevant ones through a differentiable selection, discarding the rest





1 Dataset Construction

A Triplets Creation: *{Context, Question, Answer}*

...Class Action Settlement. The Consent Decree was approved on January 8, 2020, with $2.1 million...

Short legal contexts

Teacher LLM

Question: What was the total amount paid by the county in the consent decree?
Answer: The total amount paid by the county was $2.1 million.

B **Dataset Refinement** (answer correction)

∀ Question q

Retriever

Contexts

Correct context

Keep answer generated in A

Wrong context

Substitute the answer with: "***Please provide the correct document.***"

2 Training

Correct Context Instances

Input
Context: ...The Consent Decree was finally approved on January 8, 2020, with the $2.1 million...
Question: What was the total amount paid by the county in the consent decree?

Output
Answer: The total amount paid by the county was $2.1 million.

Student SLM

Wrong Context Instances

Context: ...The event was held on March 15, 2021, celebrating the opening of...
Question: What was the total amount paid by the county in the consent decree?
Answer: Please provide the correct document.

3 Inference

Question: What was the reason for the lawsuit filed by the eight students against the Cities of Glendale and Los Angeles?

Retriever

1 ...The plaintiffs alleged that defendants...

2 ...Eight students sued the Cities of ...

Student SLM

1 Please provide the correct document.

2 **Answer:** ...students sued the cities due to the unlawful round-up of Latino students...

Selective Generative Paradigm

Right: Ace-Attorney overview. (1) A frozen LLM generates triplets from short legal contexts. For each question, we retrieve the corresponding context, manually adjusting the answer if the context is incorrect. (2) We then train an SLM on these triplets. (3) During inference, we employ SGP, where the SLM provides an answer based on the relevance of the retrieved document

Publications

- Cocchieri A., Ragazzi L., Italiani P., Tagliavini G., and Moro G. (2025). What do you call a dog that is incontrovertibly true? Dogma: Testing LLM Generalization through Humor. ACL Main
- Italiani P., Moro G., and Ragazzi L. (2025) Enhancing legal question answering with data generation and knowledge distillation from large language models. Artificial Intelligence and Law
- Ragazzi L., Italiani P., Moro G., and Panni, M. (2024). What are you token about? Differentiable perturbed top-k token selection for scientific document summarization. ACL Findings
- Italiani P., Frisoni G., Moro G., Carbonaro A., and Sartori C. (2024). Evidence, my dear watson: Abstractive dialogue summarization on learnable relevant utterances. Neurocomputing
- Moro G., Piscaglia N., Ragazzi L., and Italiani P. (2024). Multi-language transfer learning for low-resource legal case summarization. Artificial Intelligence and Law
- Frisoni G., Italiani P., Salvatori S., and Moro G. (2023). Cogito ergo summ: abstractive summarization of biomedical papers via semantic parsing graphs and consistency rewards. AAAI Main.