Poodle: Seamlessly Scaling Down LLMs with Just-in-Time Model Replacement

FG DB Spring Symposium 2026 | 4th - 5th of March | Regensburg, Germany

Nils Strassenburg, Boris Glavic, Tilmann Rabl

Just-in-Time Model Replacement

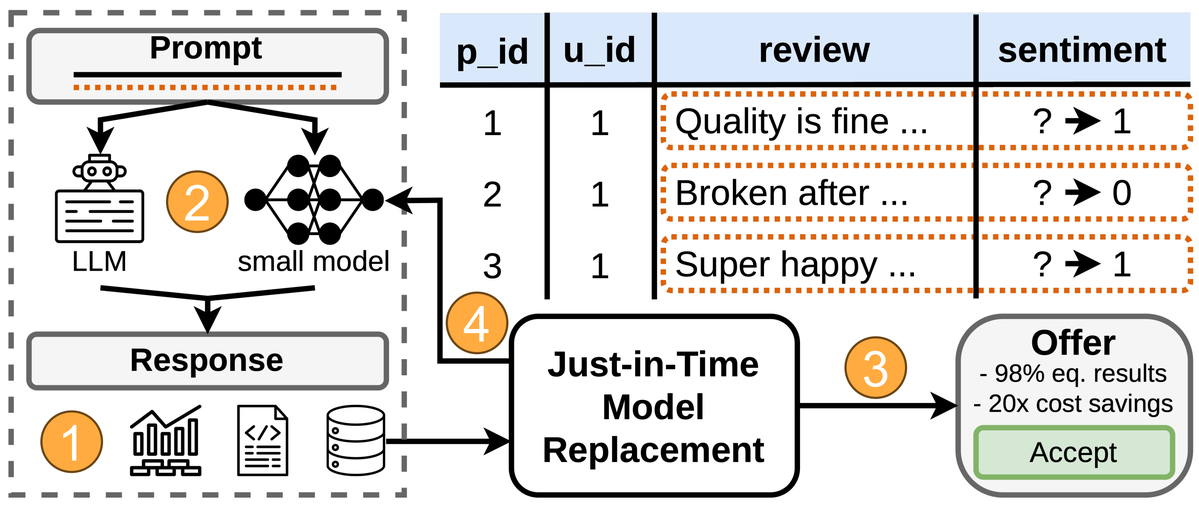
LLM

- ⊕ SotA AI capabilities
- ⊕ No in-house model development cost
- ⊖ High inference cost and resource consumption

Custom Model

- ⊖ No SotA AI capabilities
- ⊖ High in-house model development cost
- ⊕ Low inference cost and resource consumption

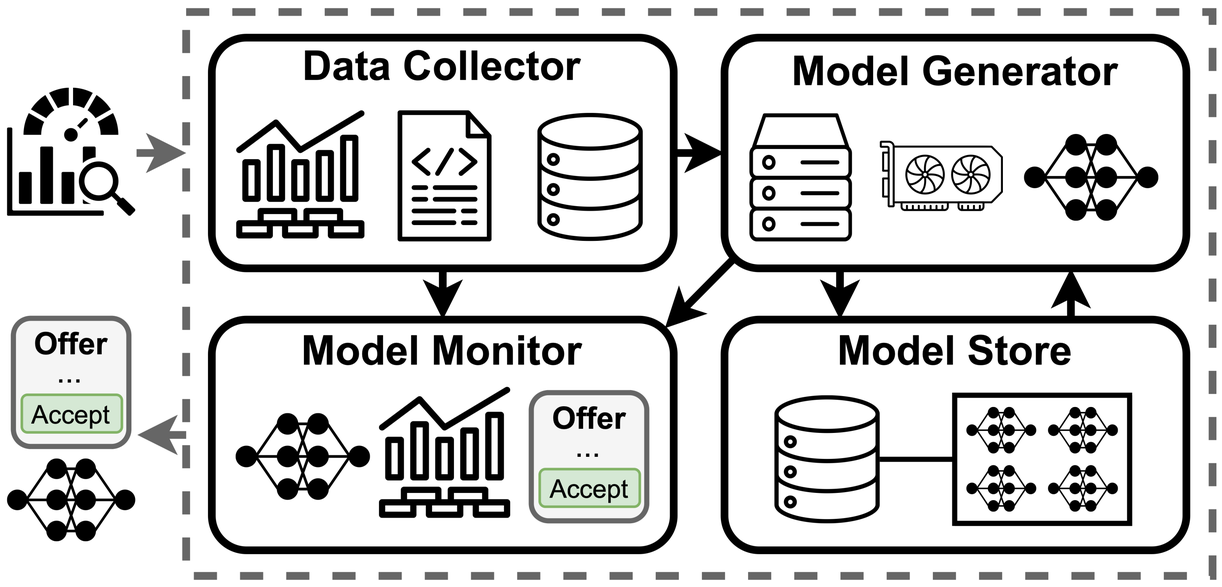
Problem: LLMs are easy to use, deliver state-of-the-art AI capabilities, and require no ML expertise. Thus, **companies offload simple reoccurring tasks to LLMs**. However, they have significantly higher costs and resource consumption than custom models.



1. Collect LLM labels, task type, task metadata
2. Develop custom model → monitor model
3. Ask user to accept switch
4. Replace LLM with custom model

Solution: Use LLM generated labels to train a custom model and **replace the LLM just-in-time with a custom model** to maintain the usability and accuracy of an LLM, but the resource consumption of a custom ML model.

Prototype and Research Directions

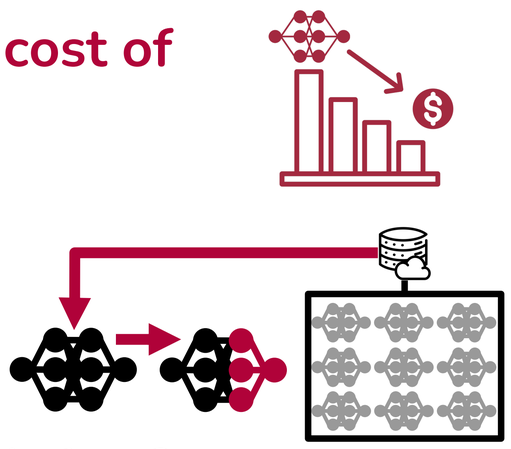


Just-in-Time Model Replacement Prototype

How can we amortize the cost of model development?

Model search → fine-tuning

- ⊕ Less training data
- ⊕ Faster convergence
- ⊕ Higher accuracy

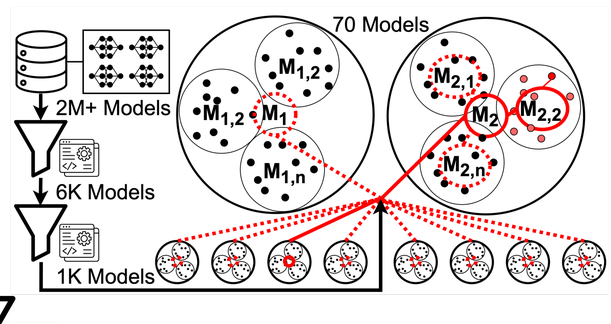


How to make model search fast?

(1) Approximate model search

(2) Co-design model stores and model search

Model Search



Metadata and indexing

Model clustering

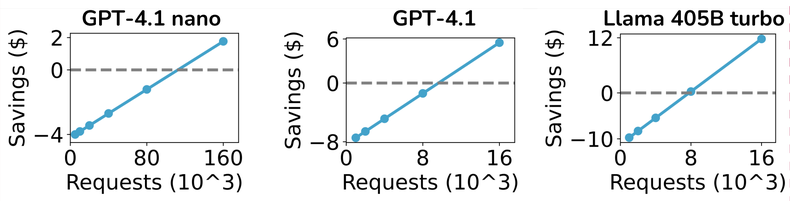
Fast model access

Multitenant optimization

Preliminary Results

JITR amortizes overhead and reduces costs significantly

Model	Input	Output	Provider
GPT-4.1	$2.00	$8.00	OpenAI
GPT-4.1-nano	$0.10	$0.40	OpenAI
Llama 405B Turbo	$3.50	$3.50	TogetherAI
Llama 8B	$0.20	$0.20	TogetherAI
BERT 80M	$0.01	$0.01	TogetherAI



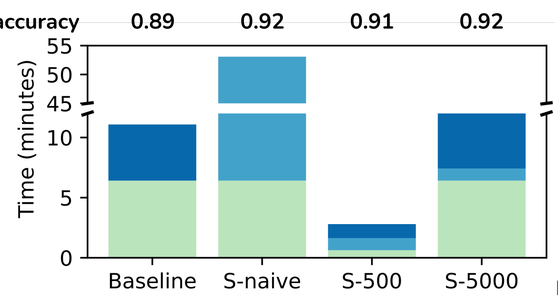
number of request until break-even

100,000 | <10,000 | <10,000

savings for 1M requests

$33 | $850 | $1420

Model search outperforms alternative approaches in dev time, accuracy, and required data



Model development time for IMDB movie review sentiment analysis.

Baseline: Fine-tune BERT

S-naïve: Exhaustive search over 10 hand-selected models from HuggingFace

S-500: Search + fine-tune on 500 items

S-5000: Search on 500 items → fine-tune on 5,000 items

Nils Strassenburg
Database Group
Hasso Plattner Institute, Potsdam, Germany
nils.strassenburg@hpi.de

This work was partially funded by the German Research Foundation (ref. 414984028 and ref. 56566056) and innovation programme (ref. 957407), and is in part supported by NSF awards IIS-2420577 and IIS-2420691.