NEURAL INFORMATION PROCESSING SYSTEMS

Large Language Model Selection with Limited Annotations

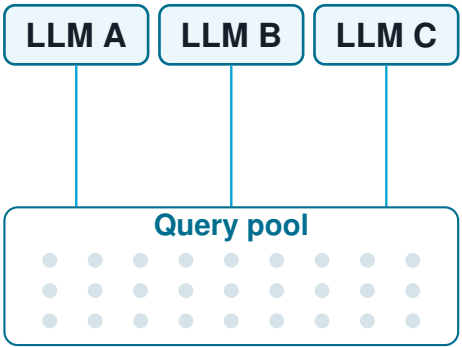
Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch,
Torsten Hoefler, Nezihe Merve Gürel

TU Delft ETH zürich

Spotlight

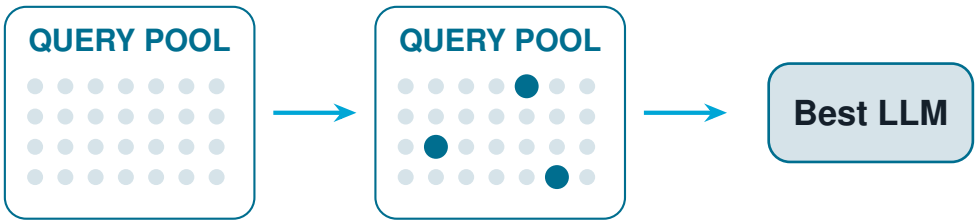
LLM Selection

- The best model depends on the application
  - Task and domain
  - Data evolves; new models emerge
- Collecting feedback for a large query pool is costly
  - Limited annotations for model selection

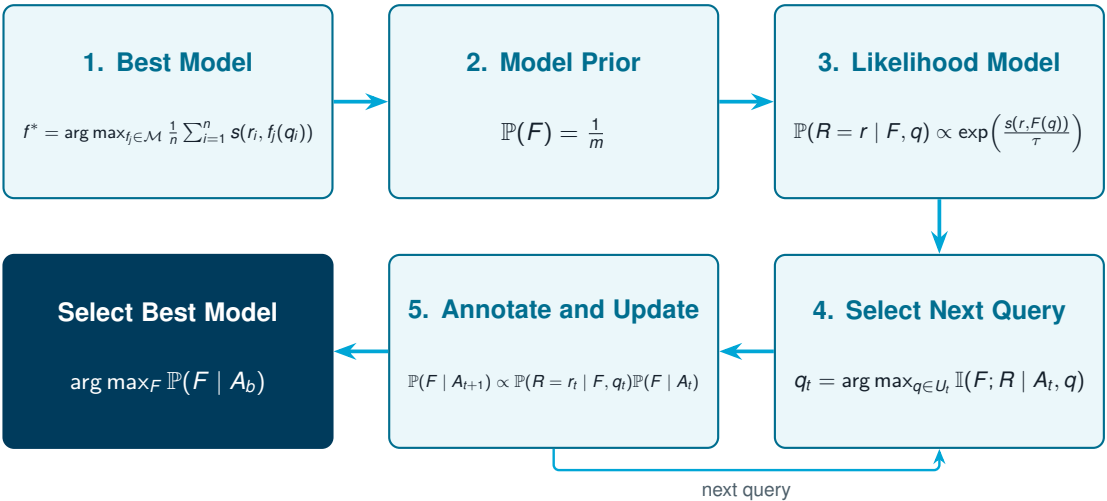


Introducing Select-LLM

- First active model selection framework for LLMs
  - Select most informative queries for annotation
  - Use annotations on selected queries to identify the best model



How Select-LLM works



How Select-LLM works

Tractable query selection rule	Posterior-weighted pairwise response similarities approximate information gain.
Model agnostic	No assumptions about model architecture or type.
Black-box	Zero access to model weights, logits, or internal states.

Evaluation at a glance

24

DATASETS

spanning reasoning, question answering, instruction following, summarization, mathematics, classification, image captioning, VQA, and translation.

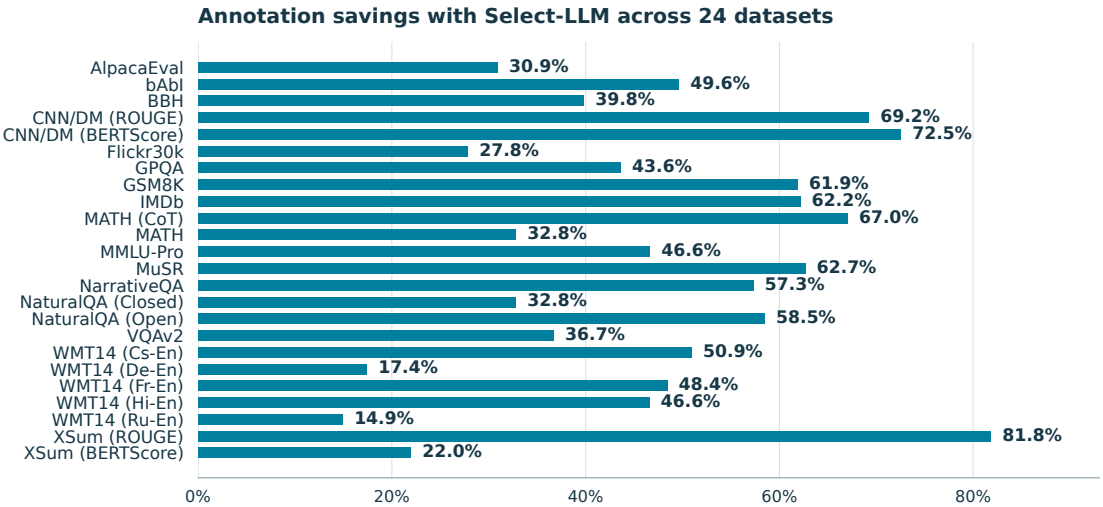
180

EVALUATED LLMs

Across diverse families and scales, spanning open-weight and API-only LLMs and multiple prompting setups.

TASK-SPECIFIC METRICS ROUGE, BLEU, BERTScore, judge preference, . . .

Find the best model with fewer labels



Fewer annotations for near-best models

- We compare annotation needs for identifying near-best models
  - Average scores in the δ-vicinity of the best model
  - Three δ thresholds.
- Select-LLM uses fewer annotations than the strongest baseline
  - In **71**/**72** cases.

Select-LLM provides a practical approach to selecting the best model in realistic black-box settings with limited annotation budgets.

Sydney Poster Session 4, Wed 9 Dec, 17:00 to 20:00 AEDT
Hall 1-4, ICC Sydney

Thank you