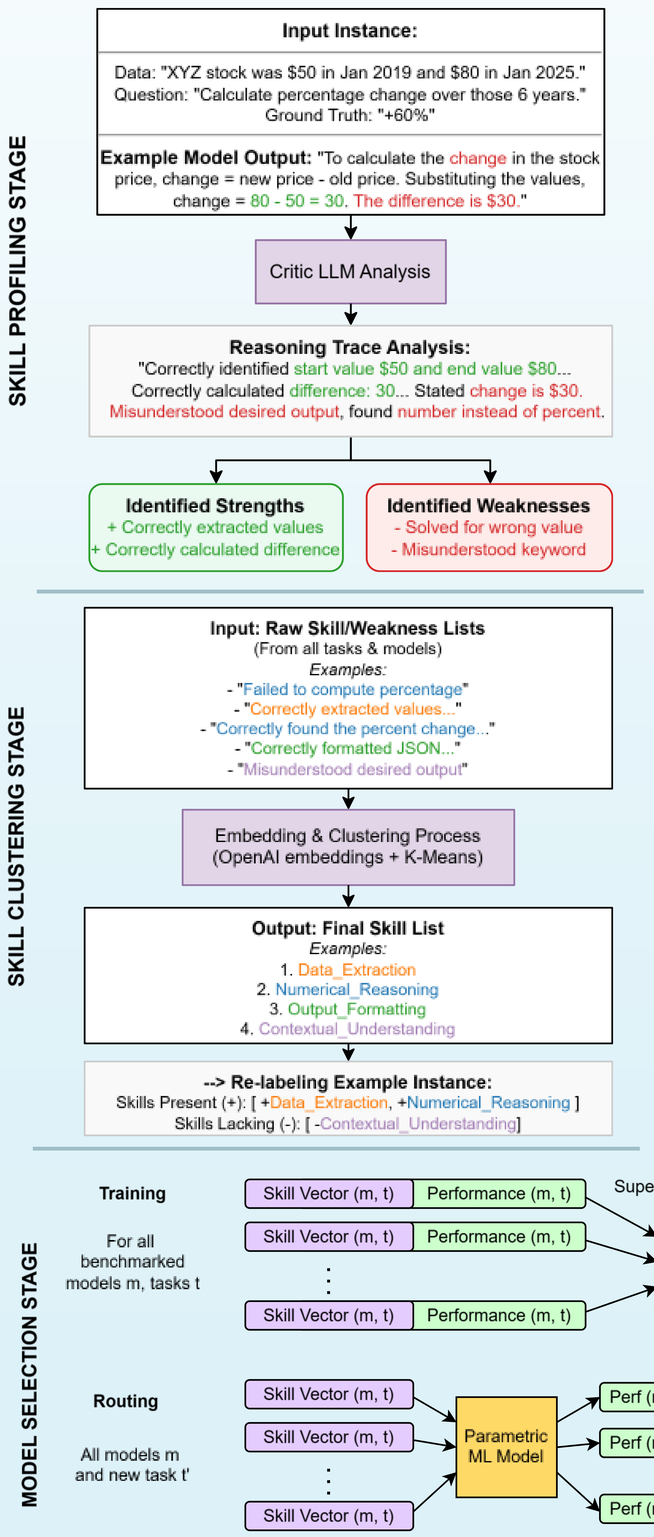
Budget-Efficient LLM Selection via Automated Skill Profiling


Mika Okamoto
Georgia Institute of Technology


Abstract

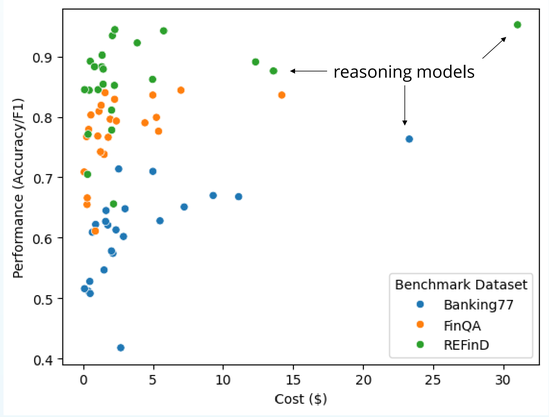
Choosing the optimal Large Language Model (LLM) presents a significant challenge due to the trade-offs between performance, cost, and energy across various models. Standard benchmarks often fail to capture the specific capabilities needed for a task or whether a less expensive model could be adequate. To address this, we introduce BELLA (Budget-Efficient LLM Selection via Automated Skill Profiling). BELLA analyzes LLM outputs to identify interpretable skills and weaknesses, creating structured profiles to recommend models that offer the best utility within user-defined resource constraints. Demonstrated initially on financial reasoning tasks, BELLA's framework is designed to be applicable across diverse domains facing cost-performance decisions.

Methodology



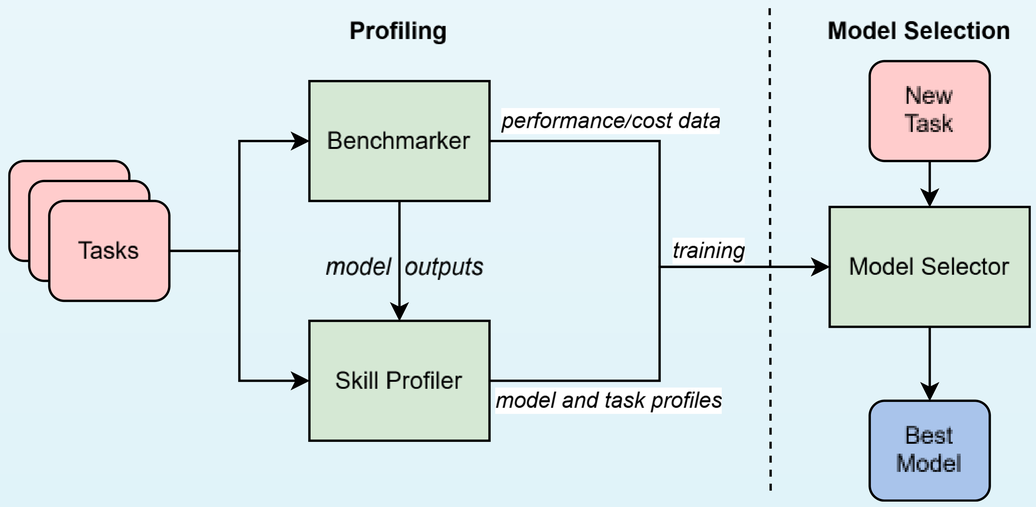
The Challenge

- **Context**: The explosion in large language model availability has created numerous options for users.
- **Challenge**: Selecting the optimal model for a task involves:
  - 1) Knowing which models perform better on your task type.
  - 2) Balancing the trade-off between performance and operational costs (compute, latency, API fees, etc).
- **Current Limitation**: Standard performance benchmarks miss:
  - 1) Skills a model needs to perform well on a given task.
  - 2) Why a model fails at a task.
  - 3) Whether a cheaper model could suffice, or what to do when the highest performing model is out of your budget.
- **Core Problem**: How to select the most suitable LLM under real-world resource constraints?



Cost/Performance visualization for example finance datasets. More costly LLMs perform better on average, but the additional performance per dollar plateaus. A cheaper model can perform just as well, for a substantial price discount.

Our Approach



- Introducing BELLA: An automated system for LLM selection.
- **Core Idea**:
  - Analyze LLM outputs on benchmark datasets to identify interpretable skills and weaknesses.
  - Build structured profiles of what models have what skills and weaknesses, and what skills a task requires.
  - Recommend models through maximizing utility within user-defined budgets (eg: cost).
- Connects granular, interpretable skill analysis directly to practical deployment constraints.

Evaluation Plan

- **Methodology**: Leave-one-task-out cross-validation on financial reasoning benchmarks.
- Comparison of selection algorithm when using features extracted from BELLA's skill profiling vs alternative methods.
- **Metrics**:
  - Cost-performance trade-off (Pareto frontier) achieved by the models selected using each feature set.
  - Agreement rate between models selected using each feature set and the optimal (Oracle) choice for each task under budget constraints.
- **Hypotheses**:
  1) Skill features derived from BELLA will enable the selection algorithm to achieve a superior cost-performance frontier compared to features from alternative profiling methods.
  2) Model selections guided by BELLA's features will exhibit higher alignment with Oracle selections than those guided by alternative features.
  3) BELLA-guided selection will effectively satisfy budget constraints while maximizing achievable performance.

Anticipated Contributions

We hope that BELLA will become a foundation for better deployment of LLMs to new tasks, and that it will bridge the gap between standardized benchmarks and real-world task constraints. Bella aims for:

- Improved Large Language Model selection for cost-effective deployment.
- Novel skill representation with analysis of model outputs to identify actual model capabilities and limitations.
- Practical resource management for new tasks building off learned skills/weaknesses of models.
- Enhanced interpretability with a transparent, capability-driven rationale for why specific models are chosen for given tasks and constraints, helping with model understanding and trust.

Future Work

- Run automated skill profiling on all LLM outputs.
- Train and evaluate model recommender according to plan.
- Evaluate BELLA across diverse domains (eg: legal, creative).
- Enable dynamic, adaptive LLM selection during tasks or for sub-components.
- Enhance multi-constraint optimization and user customization of skill importance.
- Benchmark the efficiency and overhead (added cost) of the BELLA system itself.
- Improve the explainability of BELLA's LLM recommendations.

References

Moayeri, M., Gehrmann, S., and Beutel, A. Unearthing skilllevel insights for understanding trade-offs of foundation models. arXiv preprint arXiv:2410.13826, 2024.

Zeng, Z., Wang, Y., Hajishirzi, H., and Koh, P. W. Evaltree: Profiling language model weaknesses via hierarchical capability trees. arXiv preprint arXiv:2503.08893, 2025.