Reducing Workload for Title and Abstract Screening in Medical Systematic Reviews: A Comparison of ML and LLM Approaches.

Emanuele Salonico

08.09.2025, Kick-Off Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

1. Introduction and Motivation
2. Research Questions
3. Methodology
4. Status Quo
5. Roadmap
6. References

Introduction and Motivation

Reducing Workload for Title and Abstract Screening in Medical Systematic Reviews: A Comparison of ML and LLM Approaches.

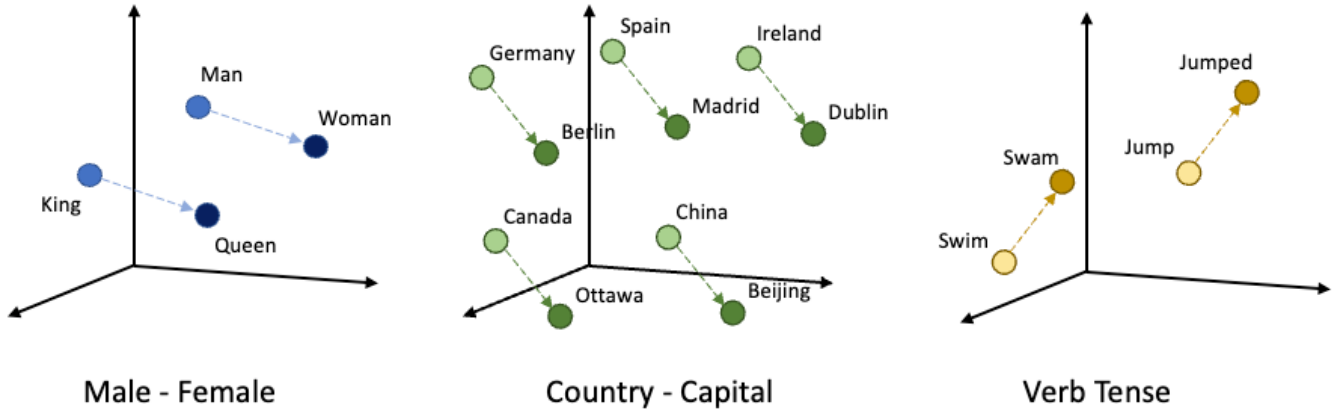
- **Goal**
  - Automate and enhance the process of paper selection in the context of systematic reviews within the medical domain.

- **Problems**
  - Manual paper selection takes time (around 4 minutes per paper)[3], and the average time from registration to publication of a systematic review is 67 weeks[1][9] .
  - Current practice usually involves at least 2 humans[1].
  - Reporting bias [1][7].
  - Amount of biomedical literature is growing, thus need for scalability.

Introduction and Motivation

- **Systematic review**
  - Structured summary of research evidence that uses predefined methods to identify, evaluate, and synthesize all relevant studies on a specific question.
- **Title and Abstract Screening (TIAB)**
  - Process of reviewing the titles and abstracts of publications to determine their relevance for inclusion in a systematic review.

- **Embeddings**
  - Dense, multi-dimensional vector representations of data (like words or documents) that capture their semantic meaning and relationships in a numerical form.



Male - Female | Country - Capital | Verb Tense

Research Questions

How can automation based on NLP and ML approaches reduce manual effort in medical literature screening for systematic reviews?

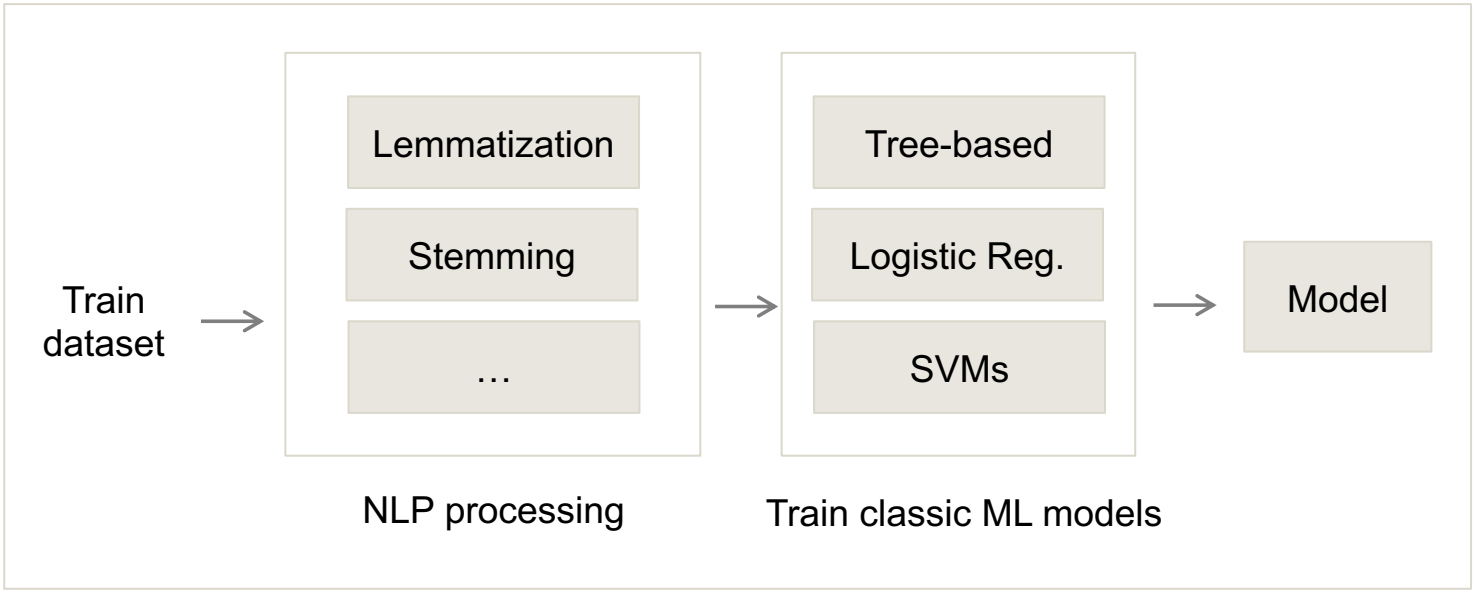
1. **RQ1:** What is the current approach in existing literature for using ML for TIAB?

2. **RQ2**: Can an embedding with a classic ML classifier pipeline alone achieve reliable include/exclude decisions in systematic reviews?
3. **RQ3**: Does end-to-end fine-tuning of a transformer meaningfully improve screening accuracy over non-tuned baselines?
4. **RQ4:** Can retrieval-augmented few-shot prompting with an LLM consistently deliver high-quality screening without any model fine-tuning?

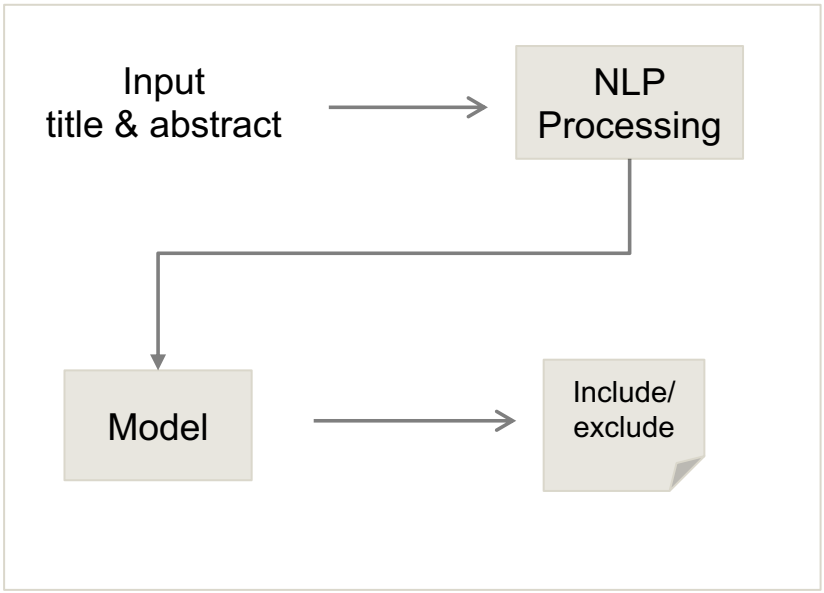
5. **RQ5**: Benchmark of the 3 proposed methods among themselves + the baseline.

Methodology - Current Approach from Literature

Data collection



Training time



Inference

Current Approach from Literature

Main findings

- Many studies report meaningful workload savings, with recall maintained or improved
- Embeddings + simple classifiers are strong, fast baselines for TIAB
- Fine-tuned transformers generally outperform keyword/rule-based baselines
- LLMs: promising for recall and reviewer burden, but variable with prompts and context

Limitations

- **Data**: Small/biased or imbalanced datasets (35% of analyzed papers)
- **Runtime/Cost**: Compute/API costs, latency (24%)
- **Prompt/Config sensitivity** (24%) in LLM workflows
- **External validity**: Generalizability/domain shift concerns (18%)
- **Reproducibility/transparency** (18%)
- **Privacy/data access** (18%)

Methodology - Proposed Approaches

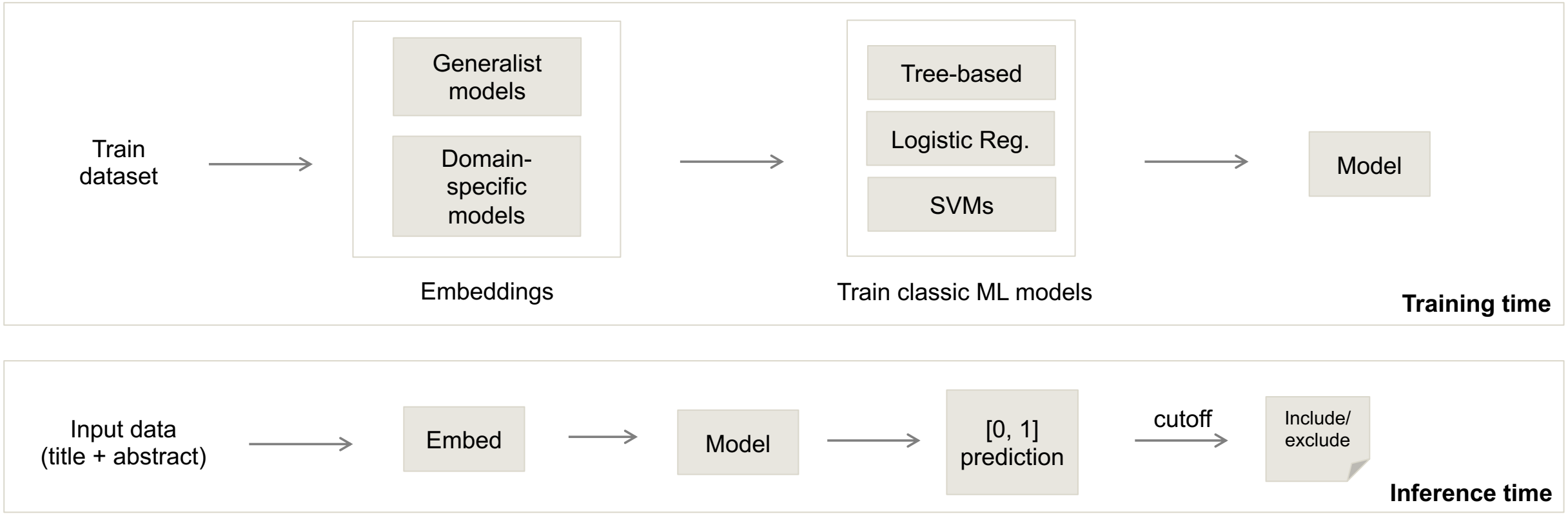
- ***Data collection*** *same as in literature (list of titles + abstract with manual annotation include/exclude).*
  - *The data scraping/collection part will not be part of the thesis. The input to the 3 methods will be an already human-labelled dataset.*

1. **Classic ML prediction on embeddings directly**
2. **Fine-tuned transformer classifier**
3. **Retrieval-augmented few-shot prompting**

(**Baseline** for comparison: manual review)

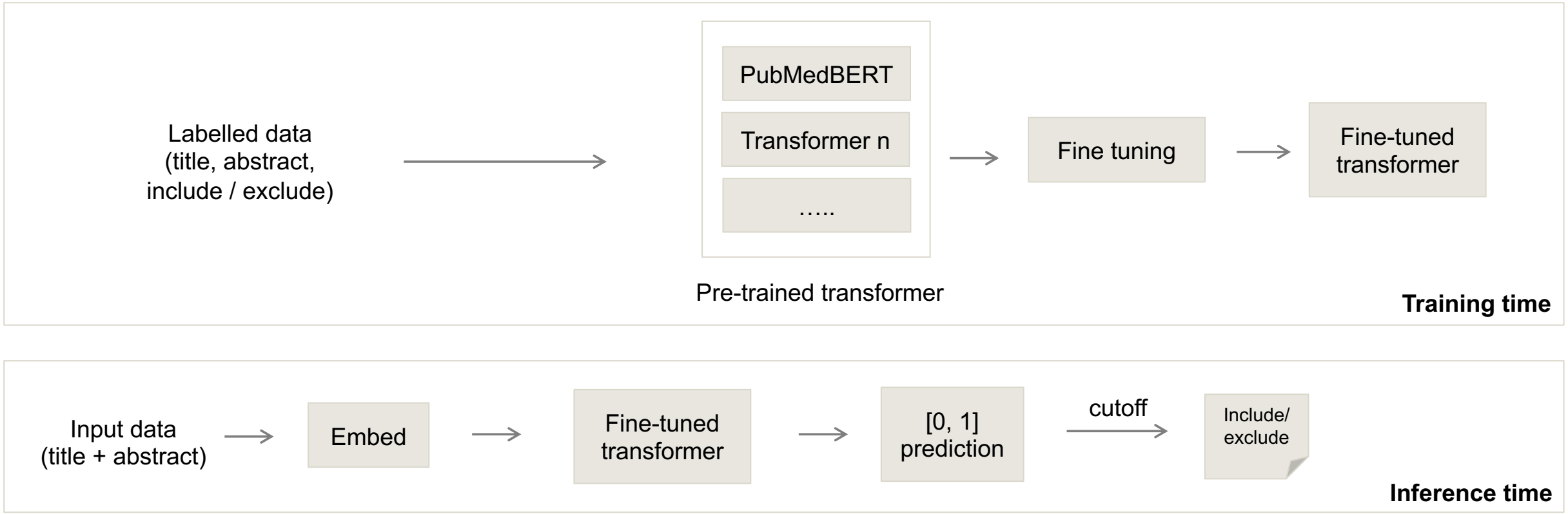
Methodology - Proposed Approaches

- **Method 1: Classic ML prediction on embeddings directly**



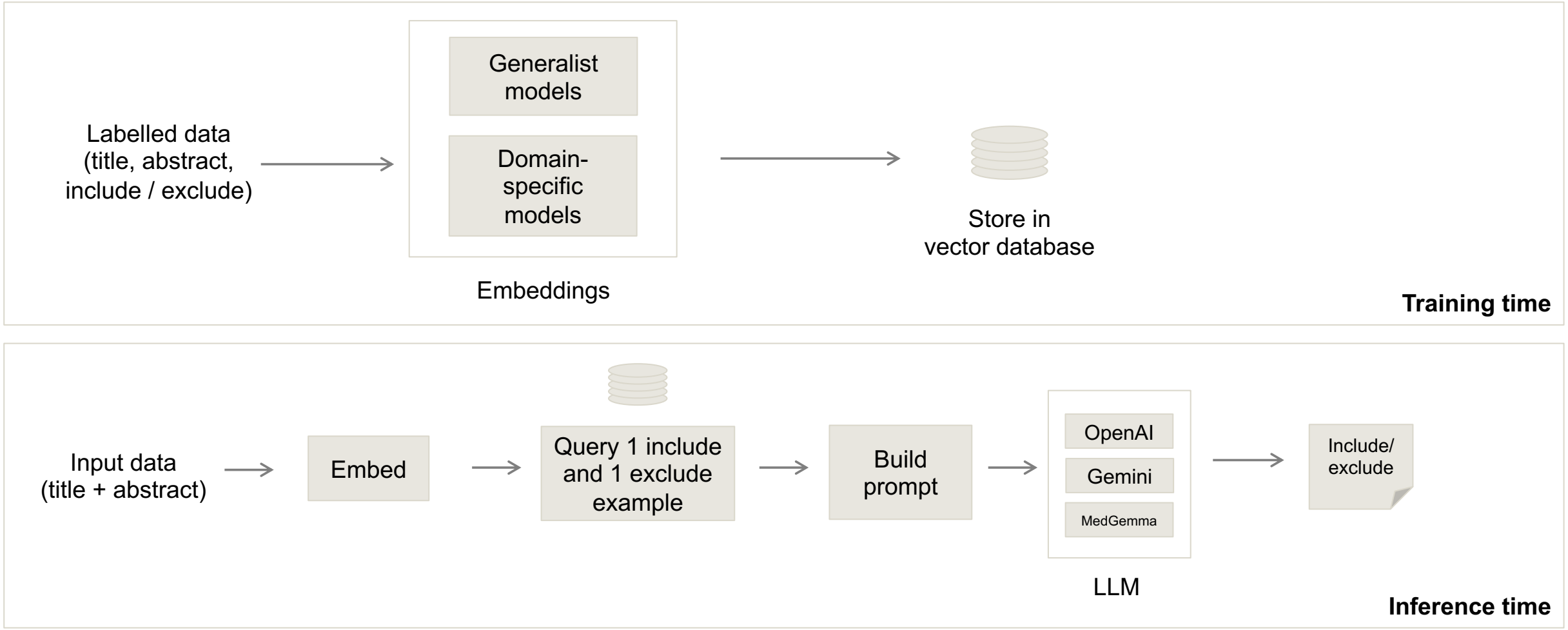
Methodology - Proposed Approaches

- **Method 2: Fine-tuned transformer classifier**



Methodology - Proposed Approaches

- **Method 3: Retrieval-augmented few-shot prompting**



Methodology - Proposed Approaches

- Throughout the whole thesis, constant **benchmarks** and **evaluations** will be carried out to come up with the most optimal, insight-based combination.

- **Embeddings**
  - Generalist models: *all-MiniLM-L6-v2, all-mpnet-base-v2, openai…*
  - Domain specific: *specter2, biomed_roberta, (Google MedGemma)…*

- **Classifier models**
  - *SVMs, logistic regression, tree-based models…*

- **Transformers**

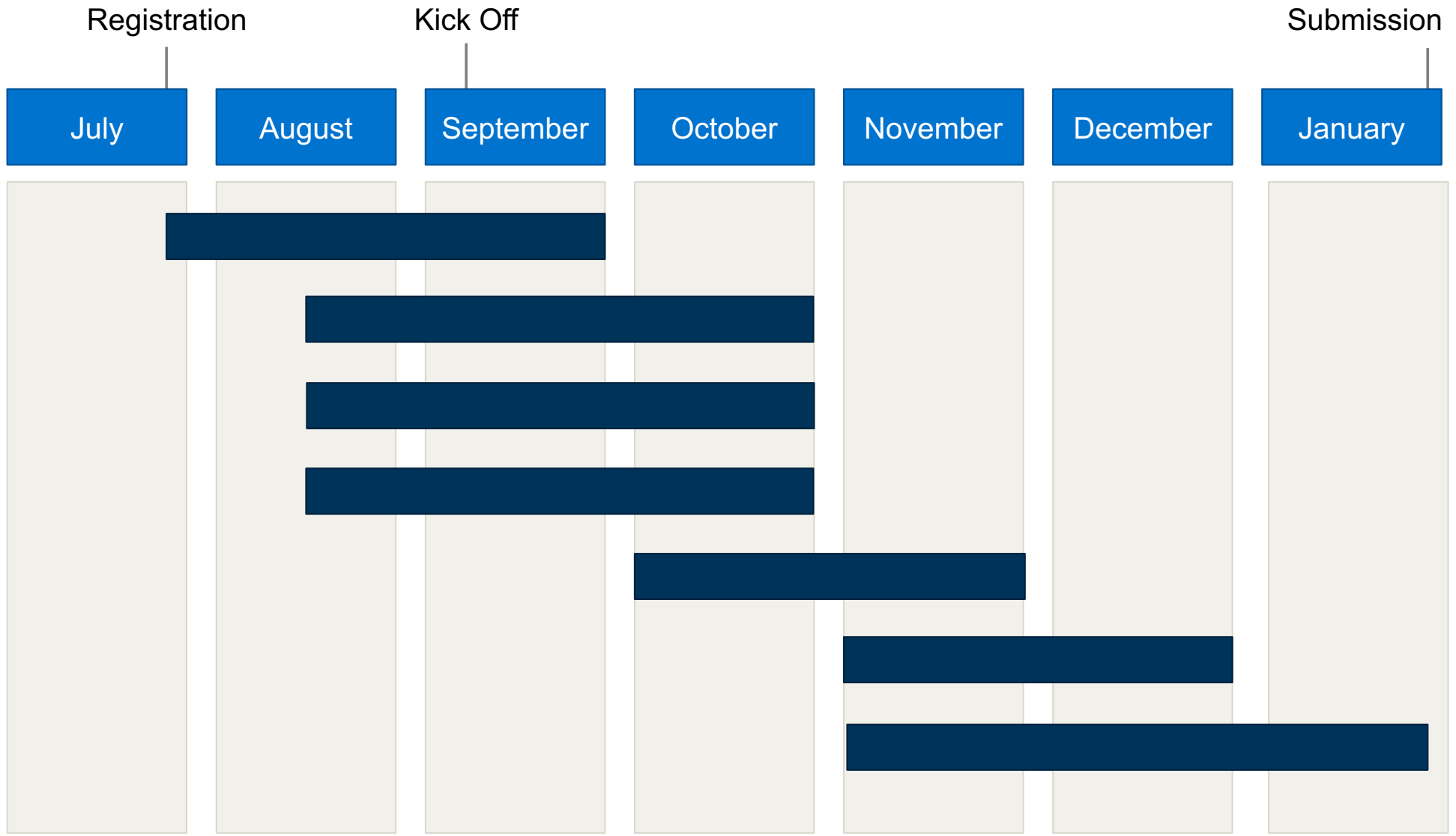
- **Vector stores**
  - *ChromaDB, FAISS…*

- **Prompting techniques**
  - *Zero-shot prompting, few-shot prompting…*

Status Quo

- **Zero-shot LLMs (GPT-3.5/4o, Llama-3)** reach ≥ 95 % recall and ~50 % workload savings in title-abstract screening[1].
- **QA-style prompting & neural rankers** (e.g., LGAR) boost precision beyond fine-tuned BERT and keyword IR baselines[1].
- **Domain-tuned embeddings** (PubMedBERT, SPECTER2) outperform TF-IDF/LSA (more basic) but demand higher compute[2].
- **Key challenges when using LLMs:** prompt variability, limited interpretability, GPU/API costs/slow routine adoption[1][9].
- **Current best practice:** human-in-the-loop via existing systematic review tools, with domain-specific validation before deployment[8].

Roadmap



References

1. Akinseloyin, O., Jiang, X. and Palade, V. (2024) '**A question-answering framework for automated abstract screening using large language models**', *Journal of the American Medical Informatics Association*, 31(9), pp. 1939–1952. doi:10.1093/jamia/ocae166.

2. Gu, Y. *et al.* (2021) '**Domain-specific language model pretraining for biomedical natural language processing**', *ACM Transactions on Computing for Healthcare*, 3(1), pp. 1–23. doi:10.1145/3458754.

3. Serrato-Fonseca, J.I. (no date) '**ACAS: A comprehensive framework for automatic abstract screening in systematic literature reviews**', Les Cahiers du GERAD [Preprint], (September 2024).

4. Hasny, M. *et al.* (2023) '**Bert for complex systematic review screening to support the future of medical research**', *Lecture Notes in Computer Science*, pp. 173–182. doi:10.1007/978-3-031-34344-5_21.

5. Kebede, M.M., Le Cornet, C. and Fortner, R.T. (2022) '**In-depth evaluation of Machine Learning Methods for Semi-automating article screening in a systematic review of mechanistic literature**', *Research Synthesis Methods*, 14(2), pp. 156–172. doi:10.1002/jrsm.1589.

6. Lange, T. *et al.* (2021) '**Machine learning for identifying relevant publications in updates of systematic reviews of Diagnostic Test Studies**', *Research Synthesis Methods*, 12(4), pp. 506–515. doi:10.1002/jrsm.1486.

7. Nykvist, B. *et al.* (2025) '**Testing the utility of GPT for title and abstract screening in environmental systematic evidence synthesis**', *Environmental Evidence*, 14(1). doi:10.1186/s13750-025-00360-x.

8. Sanghera, R. **High-performance automated abstract screening with large language model ensembles**.

9. Pilz, M. *et al.* (2024) '**Semi-automated title-abstract screening using Natural Language Processing and machine learning**', *Systematic Reviews*, 13(1). doi:10.1186/s13643-024-02688-w.