

Developing a scalable pipeline for data extraction from clinical letters through resource-efficient prompt engineering

Ariel Yuhan Ong,^{1,2} Quang Nguyen,² Ishani Barai,¹ Gregory Maniatopoulos,³ Yemisi Takwoingi,⁴ Alastair K Denniston,⁴ Pearse A Keane^{1,2}

1: Moorfields Eye Hospital NHS Foundation Trust; 2: Institute of Ophthalmology, University College London; 3: Centre for Healthcare Innovation, Policy and Management, University of Leicester; 4: Department of Applied Health Sciences, University of Birmingham, Birmingham

BACKGROUND

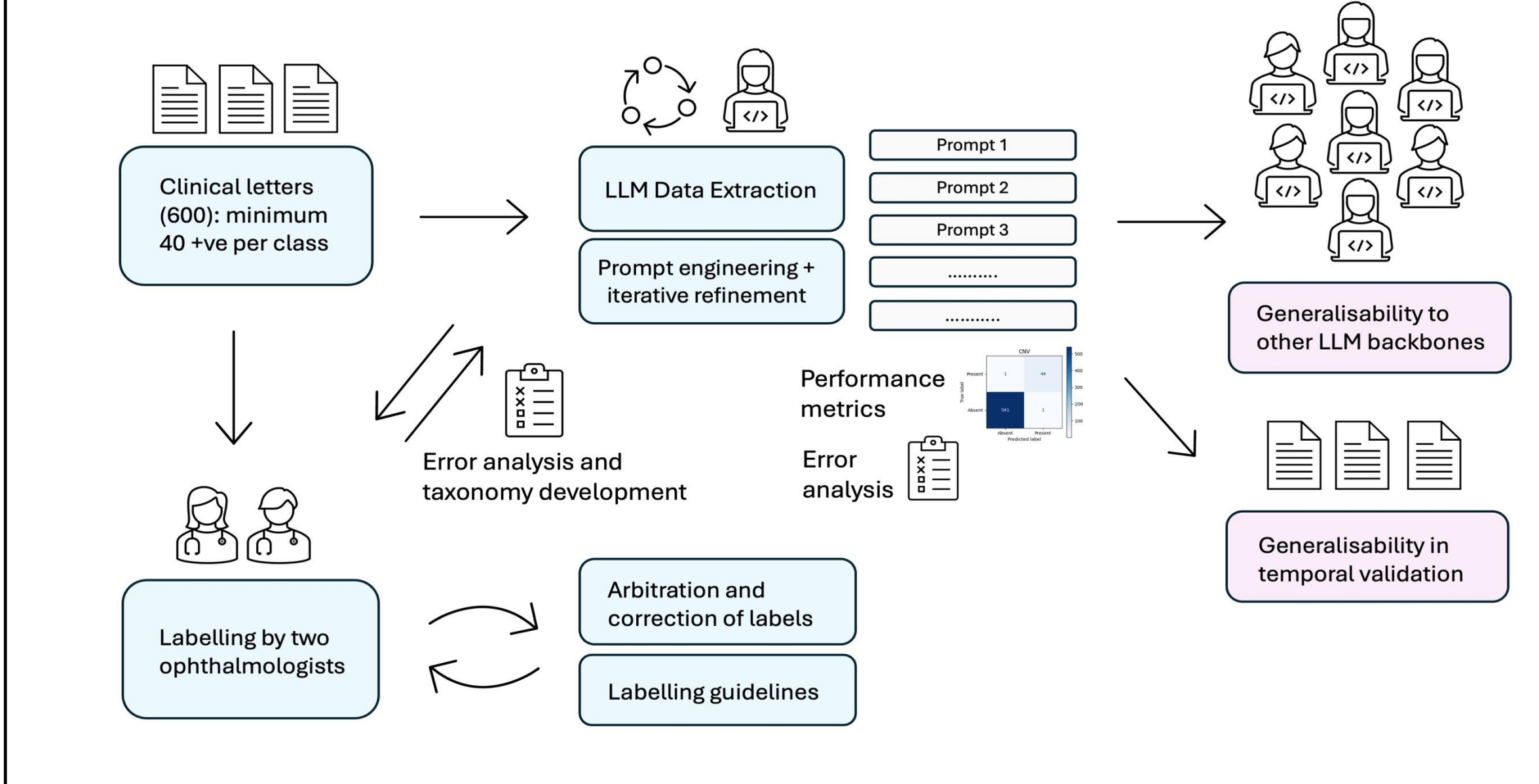
Although electronic health records (EHRs) have revolutionised healthcare delivery, secondary use for research presents a misalignment due to data quality and availability. ~80% of EHR data exists in an unstructured format (e.g. as clinical letters, discharge summaries, progress notes) and are poorly accessible for downstream research. Manual data extraction remains a significant bottleneck due to resource requirements and risk of human errors.

We present a **scalable approach** to maximise the utility of this valuable resource. As a proof-of-concept, we apply our data extraction pipeline to extract **9 common macular diseases and laterality** from ophthalmic clinical letters, to generate reference labels for evaluating a deep learning algorithm.

OBJECTIVES: To develop a scalable pipeline for the complex task of extracting information from real-world ophthalmic clinical letters

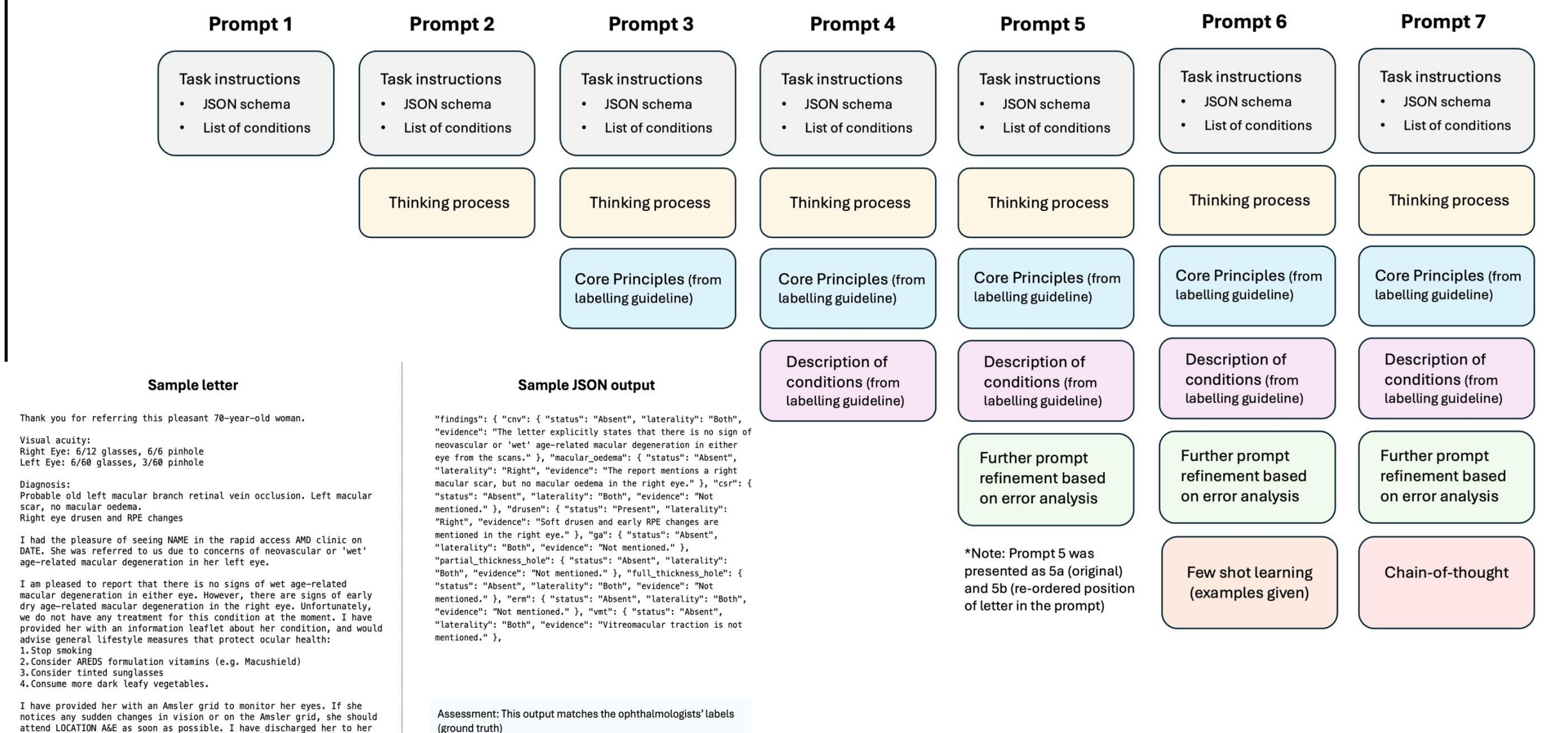
METHODS

Study workflow including 1) labelling of clinical letters, 2) prompt engineering and iterative refinement using a development model, 3) extending this approach to other LLM families, and 4) testing generalizability in temporal validation



Approach to iterative prompt refinement

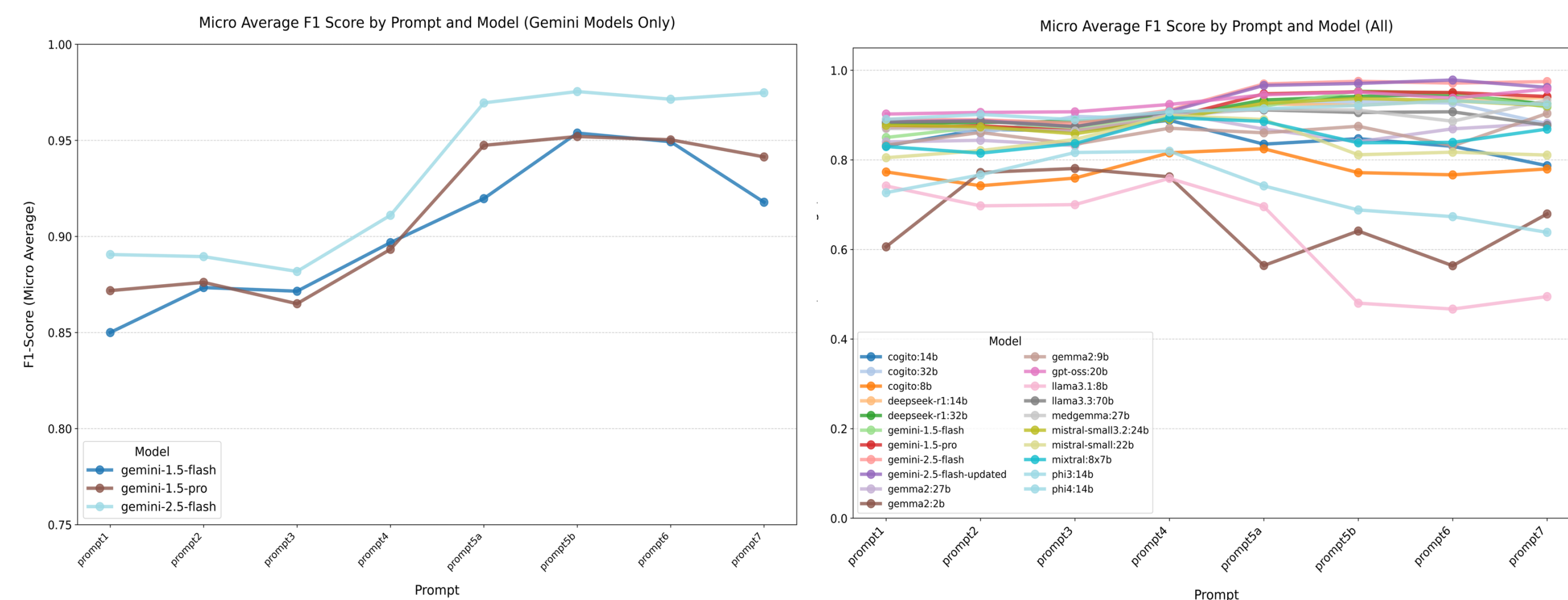
We present a modular framework to prompt engineering – prompts 1 to 4 involve the addition of new modular instructions, prompt 5 incorporates further refinement from error analyses of failure modes, and prompt 6 and 7 incorporate advanced prompt engineering techniques



RESULTS

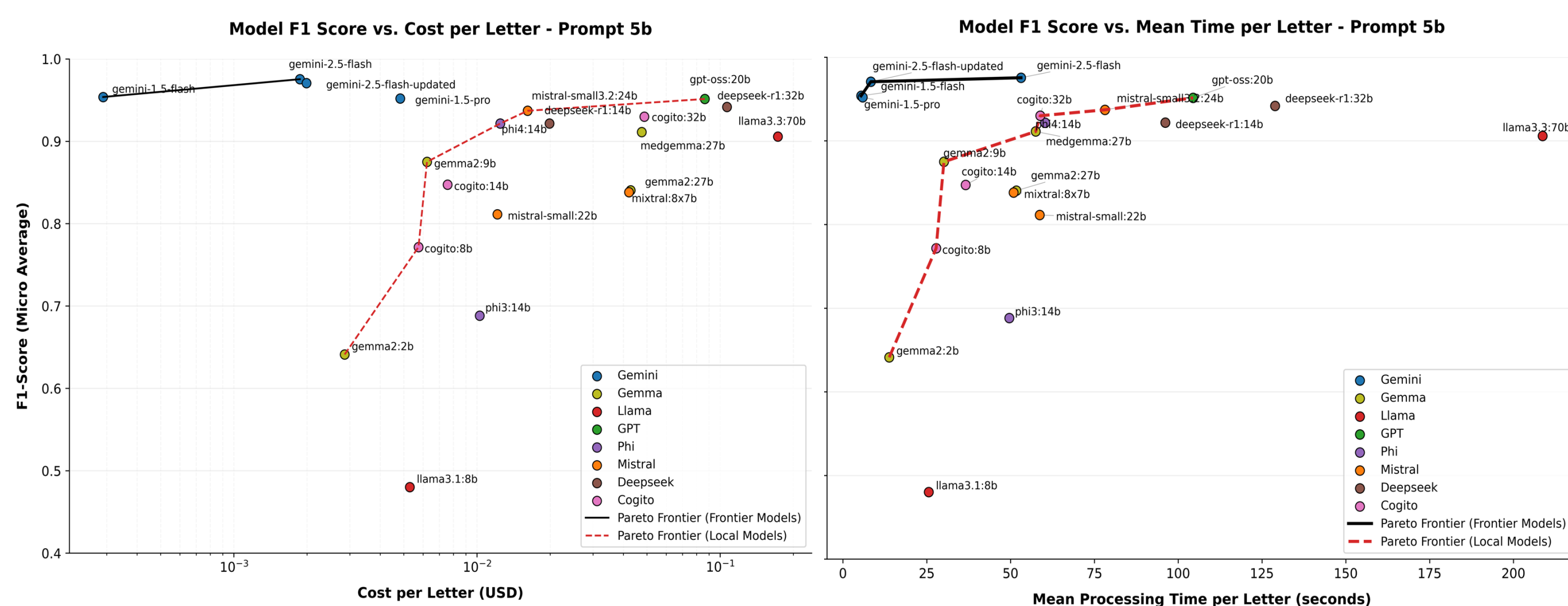
Our pipeline achieved strong performance with a proprietary model in development, yielding a **maximum micro-averaged F1 score of 0.954 (95% CI 0.941-0.967)**

This approach **extended to other models in the same family** and **17 LLMs from 7 open-weight LLM families**.



We adapt the **“Pareto Frontier”**, a concept borrowed from economics, to consider the trade-off between two objectives (cost vs performance, time vs performance) for scaling up

This approach allowed us to select Gemini 2.5 Flash (with prompt 5b) as the optimal model-prompt configuration for scaling up to 94,904 letters



We present an inductively-developed **error taxonomy** comprising 7 themes categorised as 1) model-centric errors and 2) data-centric errors to understand model performance and inform future refinement. We calibrate and apply an “LLM judge” to classify all errors across all model-prompt configurations



TAKE HOME MESSAGES

- We have developed a **high-performing LLM-based data extraction pipeline** which demonstrates generalisability (achieved micro-F1 0.945-0.980 in temporal validation)
- We present a **multi-dimensional evaluation framework for deployment**, including an error taxonomy and Pareto frontier analyses to systematically map the operational trade-offs across different LLM configurations.
- A robust approach to operationalisation in real-world workflows at scale may help lay the foundation for next-generation data pipelines that accelerate scientific discovery and power continuous learning health systems

Future work: extension across tasks and other specialties is ongoing; uncertainty quantification with audit mechanisms are planned to provide reassurance for real-world deployment at scale

Preprint available at:

