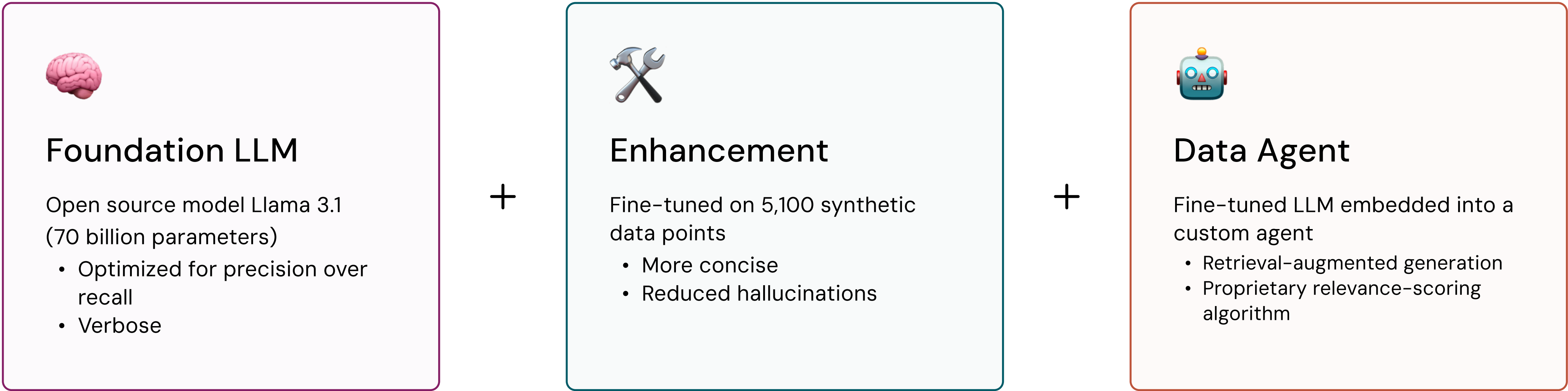


Introduction

SLRs are incredibly labour-intensive and expensive, costing pharmaceutical companies approximately \$141,194.80 per SLR, and up to \$16M per year¹. Large language models (LLMs), a type of AI capable of reading and processing text, offer great potential to reduce costs and accelerate timelines. Previous research has examined the ability of LLMs to match human reviewers in the screening step of SLRs. However, off-the-shelf foundation LLMs, such as ChatGPT, are not tailored for this use case – not optimized for recall, they leave out key insights. To investigate this, we fine-tuned an open-source LLM (Llama 3.1) for an SLR workflow, applied it to our proprietary literature review software (Reliant Tabular), and measured its performance against a recent benchmark study² comparing GPT-4 and human analyst effectiveness for SLR workflows.

Methods

Reliant Tabular System Architecture



Model

Our software acts as a data agent that reads through each abstract to perform a high-recall screening step. Fine-tuned from Llama 3.1, it focuses on short, correct outputs with minimal hallucinations, and outputs “No answer” when an answer cannot be determined.

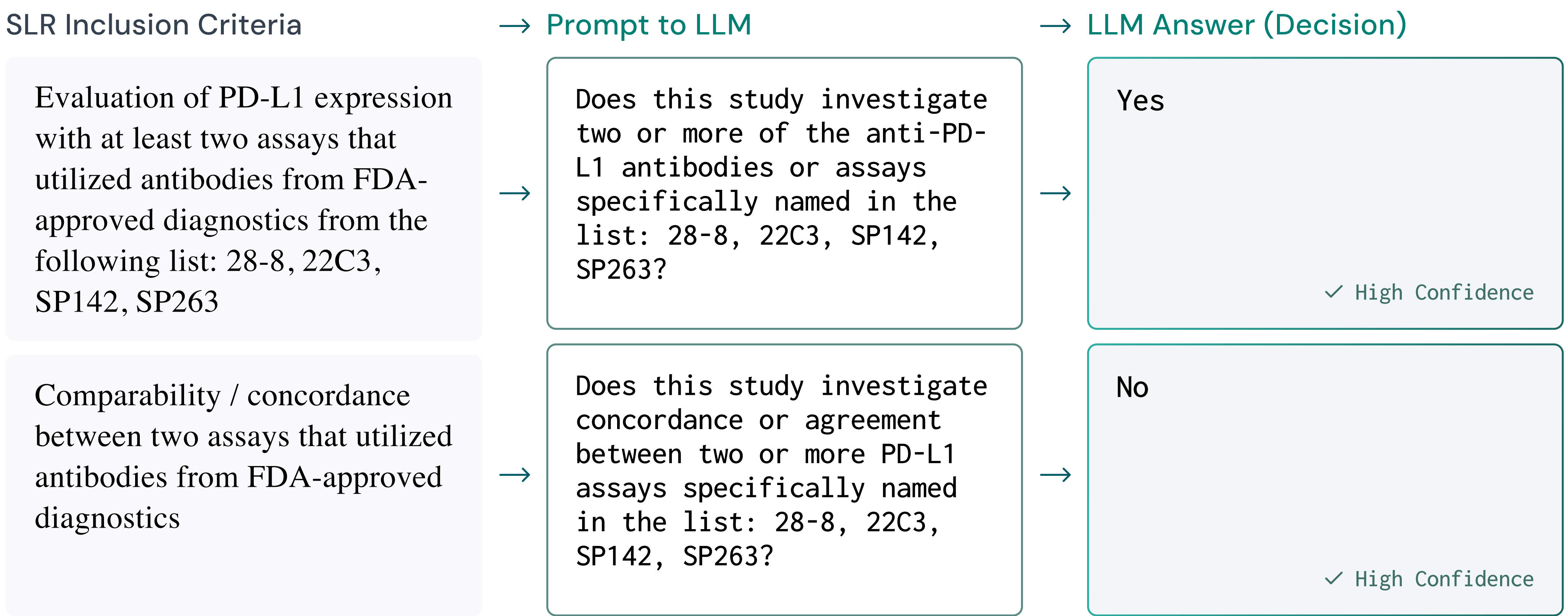
Fine-tuning is performed using a synthetic dataset of 5,100 examples created through a 'reverse extraction' process—starting with an answered query and generating a plausible user scenario that could have prompted it. Fine-tuning not only manages costs, but also produces faster outputs by controlling the length of the model’s answers.

Document set

982 English-language MEDLINE abstracts of varying complexity, available on PubMed. Sourced from “Analytical Concordance of PD-L1 Assays Utilizing Antibodies From FDA-Approved Diagnostics in Advanced Cancers: A Systematic Literature Review”³



SLR Replication

We used our software, powered by the enhanced LLM, to perform a screening step on the set of 982 abstracts. For each abstract, it applied the screening prompts from Kerr et al., generated a response, validated it, and provided confidence signals.



Evaluation

We compared our data agent to GPT-4 and a human analyst against the ground truth number of 55 abstracts determined to to be relevant (“positives”).

	 GPT-4	 Reliant Model
False Positive Rate	8.5%	5.3%
False Negative Rate	11.3%	0.0%

*using mean of 5 runs; FP 52 ± 5.1, FN 0.2 ± 0.4

Findings

- + Of the original 982 abstracts, our data agent identified 107 abstracts for further review, capturing all 55 abstracts that human reviewers marked positive for inclusion, demonstrating an effective recall rate of 100% on a real world workflow.
- + Our system significantly outperformed GPT-4 in recall, producing fewer false positives and false negatives.
- + At 0.006 minutes per abstract, our system was 833x faster than a human at 5 minutes per abstract – a total of nearly 82 hours saved.

High recall is paramount for SLRs: while false positives can be further reviewed and excluded, false negatives lead to entirely missed insights, requiring extensive rework to recover. The Reliant fine-tuned model minimizes this critical error type, outperforming GPT-4 on systematic literature reviews, in just a fraction of the time.

¹ Michelson M., Reuter K., Contemp Clin Trials Commun 2021;16:100443, “The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials”.

² Kerr B., et al. ISMPP US 2024, “Concordance between generative artificial intelligence and human reviewers for screening of publications for a systematic literature review”.

³ Prince EA., et al. JCO Precis Oncol. 2021;5:953-973, “Analytical Concordance of PD-L1 Assays Utilizing Antibodies From FDA-Approved Diagnostics in Advanced Cancers: A Systematic Literature Review”.

⁴ Feng Y., et al. J Am Med Inform Assoc. 2022;29(8):1425-1432, “Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis.”