Reproducing and extending "Look-Ahead Planning in Large Language Models"

Alexander Holmberg (alholmbe@kth.se)
Pontus Linde (ponlin@kth.se)
Dante Wesslund (dantew@kth.se)

October 2025

1 Topic and Paper

This project falls under the topic **Understanding Deep Networks**, with a focus on **mechanistic interpretability**. We will reproduce and extend the paper *"Look-Ahead Planning in Large Language Models"* [1]. The paper investigates how large language models internally represent planning behavior through information flow and probing analyses. Given that no public implementation exists and that the work requires building all experiments from scratch, we target the **Excellent (A–B)** grade category.

2 Project Plan

2.1 Overview

The goal of the project is to reproduce the interpretability analyses presented in the original paper and evaluate whether similar look-ahead mechanisms appear in current open-source models. Specifically, we will test whether internal representations of transformer layers encode both *current states* and *future decisions* in planning tasks. Our research questions are:

- Can the main findings on information flow and internal representations be reproduced on the original models (LLaMA-2-7B and Vicuna-7B)?
- Extension 1: Do open-weight models that are state-of-the-art today in late 2025 (e.g. Qwen 3.1-7B) exhibit stronger or deeper look-ahead planning?
- Extension 2: How does planning depth evolve when the number of reasoning steps is increased?

2.2 Experiments

- Reimplement the extraction-rate analysis for MLP and MHSA layers to determine where decision-relevant information emerges within the network.
- Reproduce the information-flow computation between goal, history, and action tokens using attention-gradient products to measure how information propagates across layers.
- Train linear and nonlinear probes on hidden representations to evaluate how well different layers encode both the current block configuration and upcoming decisions.
- **Extension 1**: Repeat all or some of the experiments on modern open-weight models (e.g., Qwen 3.1–7B).
- **Extension 2:** Increase the planning horizon in the Blocksworld dataset (from 6 to 8–10 steps).

2.3 Datasets

We will re-create the synthetic Blocksworld dataset described in the paper. Each instance consists of textual descriptions of the initial and goal states of colored blocks and a sequence of optimal actions ("pick up red", "stack red on blue", etc.). Following the paper, we will generate between 4 and 6 colors, 4 piles, and up to 6 steps per plan. In later experiments with more capable models, we will extend the dataset to 8–10 steps to study longer-horizon planning.

3 Justification for Aimed Grade

We aim for the **Excellent (A–B)** category. The project requires full reimplementation of all components, as no public code is available for the target paper. This includes recreating the synthetic Blocksworld data generator, model fine-tuning, and all interpretability analyses (extraction-rate, information-flow, and probing). This alone make this paper harder than others in some sense. Beyond reproduction, we will conduct two principled extensions: (i) repeat the analyses on a modern open-weight 7B model such as Qwen 3.1–7B, and (ii) increase the planning horizon from 6 to 8–10 steps to test whether look-ahead reasoning scales with task complexity.

4 Computational Requirements and Resources

The original paper reports full fine-tuning of 7B models on Blocksworld. We will start with parameter-efficient fine-tuning (LoRA/QLoRA) for feasibility, which fits comfortably on a single 24–40 GB GPU. If resources allow, we might also perform full fine-tuning runs for comparison. We have access to a 80GB A100 GPU that we will use. We will use PyTorch with the Hugging Face Transformers and PEFT libraries for fine-tuning (LoRA/QLoRA). Additional packages include bitsandbytes for quantization, matplotlib/seaborn for visualization, and standard Python tooling for dataset generation and probe training.

5 Evaluation and Success Criteria

Success will be measured by the extent to which our reproduced analyses match the qualitative and quantitative trends reported in the paper. We expect similar extraction-rate curves, comparable information-flow patterns, and probe F1-scores that reflect both current and future state encoding.

6 Expected Learning Outcomes

Alexander Holmberg: I want to learn more about mechanistic interpretability, and how to do conduct thorough analysis of how the internals of LLMs work. **Pontus Linde:** I want to learn the internals of LLMs along with gaining experience of how to fine-tune efficiently with LoRA. **Dante Wesslund:** I'm interesting in learning how to fine-tune models on GPUs, and to compare the performances in the older models with today's state of the art.

References

[1] Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models, 2024.