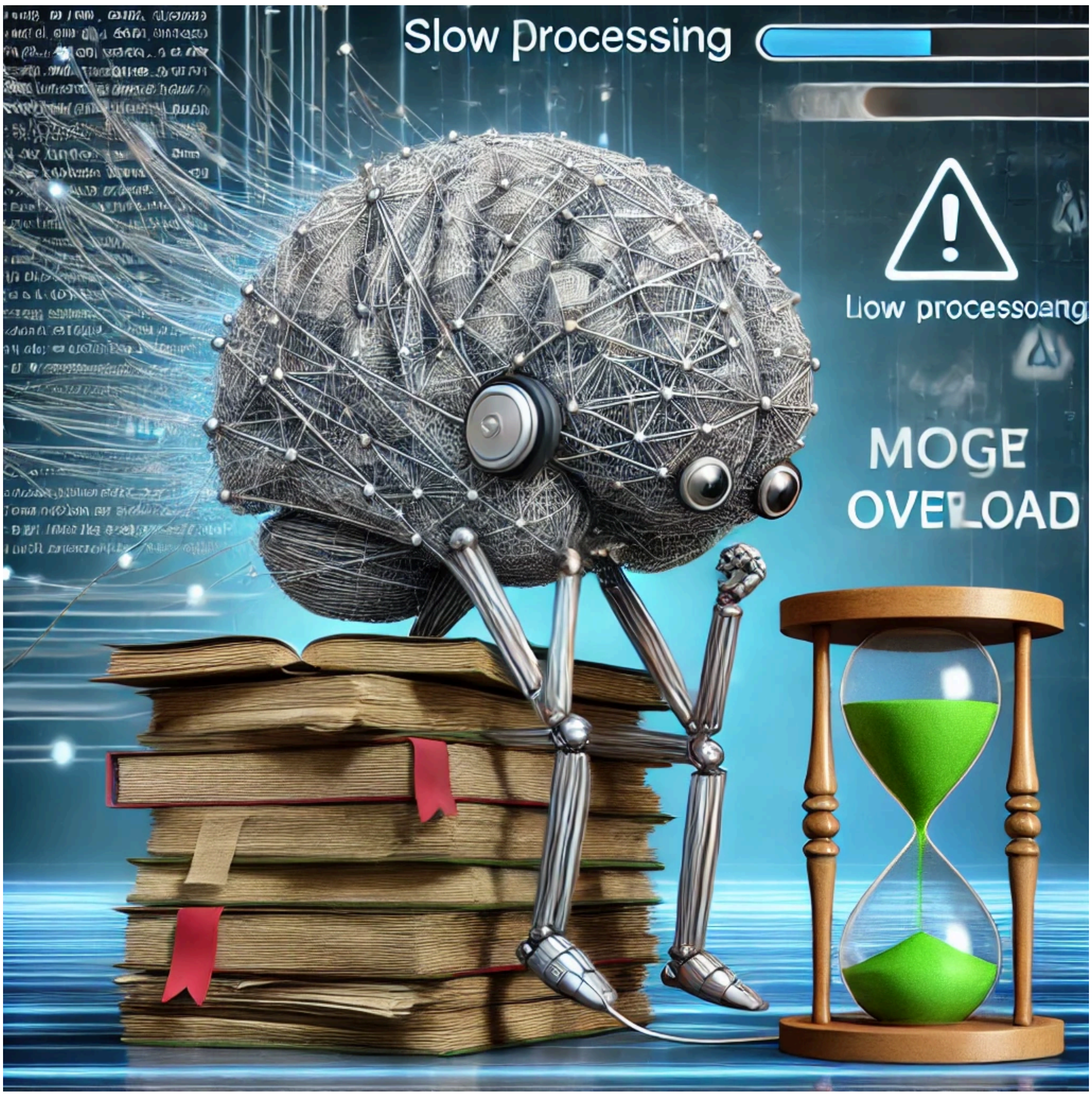
🚀 Mixture of Block Attention (MoBA): A Smarter Way to Handle Long Texts

🔍 The Problem: Why Do AI Models Struggle with Long Texts?



💡 The Solution: What Is MoBA?

Mixture of Block Attention (MoBA) is a new approach that **makes attention faster and more efficient**. It applies the principles of **Mixture of Experts (MoE)** to attention, meaning the model **focuses only on the most important parts of the text** instead of processing everything equally.

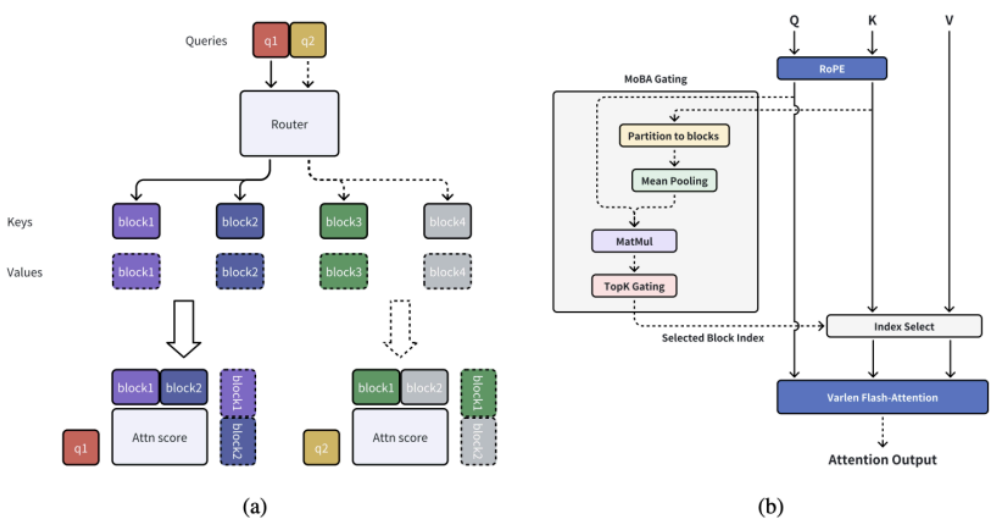


Figure 1: Illustration of mixture of block attention (MoBA). **(a)** A running example of MoBA; **(b)** Integration of MoBA into Flash Attention.

🔹 How Does MoBA Work?

1️⃣ **Breaks the text into blocks** instead of looking at every token individually.
2️⃣ **Uses a trainable gating system** to decide which blocks are most relevant.
3️⃣ **Allows flexible attention**, meaning it can still focus on distant but important information.
4️⃣ **Maintains standard Transformer properties**, so it works with existing AI models.

Benchmark	Llama-8B-1M-MoBA	Llama-8B-1M-Full
AGIEval [0-shot]	0.5144	**0.5146**
BBH [3-shot]	0.6573	**0.6589**
CEval [5-shot]	**0.6273**	0.6165
GSM8K [5-shot]	**0.7278**	0.7142
HellaSWAG [0-shot]	0.8262	**0.8279**
Loogle [0-shot]	**0.4209**	0.4016
Competition Math [0-shot]	0.4254	**0.4324**
MBPP [3-shot]	**0.5380**	0.5320
MBPP Sanitized [0-shot]	**0.6926**	0.6615
MMLU [0-shot]	0.4903	**0.4904**
MMLU Pro [5-shot][CoT]	0.4295	**0.4328**
OpenAI HumanEval [0-shot][pass@1]	0.6951	**0.7012**
SimpleQA [0-shot]	0.0465	**0.0492**
TriviaQA [0-shot]	**0.5673**	0.5667
LongBench @32K [0-shot]	**0.4828**	0.4821
RULER @128K [0-shot]	0.7818	**0.7849**

Table 2: Performance comparison between MoBA and full Attention across different evaluation benchmarks.

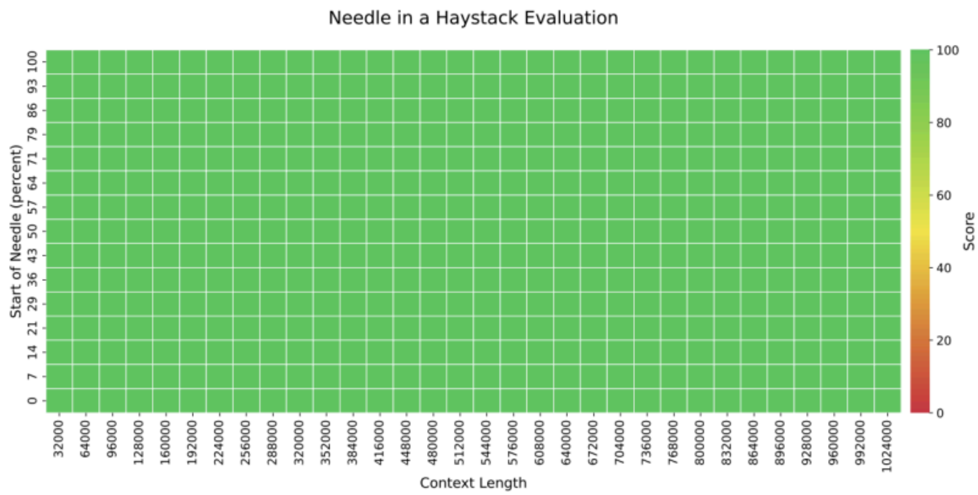


Figure 7: Performance of LLama-8B-1M-MoBA on the Needle in the Haystack benchmark (upto 1M context length).

💪 POPPER: Automated Hypothesis Validation with Agentic Sequential Falsifications

🔍 The Importance of Hypothesis Validation

Hypothesis validation is fundamental in **scientific discovery, decision-making, and information acquisition**. Whether in **biology, economics, or policymaking**, researchers rely on testing hypotheses to guide their conclusions. Traditionally, this process involves:

- Designing experiments
- Collecting data
- Analyzing results

However, **Large Language Models (LLMs)** have dramatically increased the number of generated hypotheses. While these AI-driven insights offer potential breakthroughs, their **plausibility varies widely**, making **manual validation impractical**. Automating this process is essential to ensure only **scientifically rigorous hypotheses** guide future research.

💡 POPPER: AI-Driven Hypothesis Validation

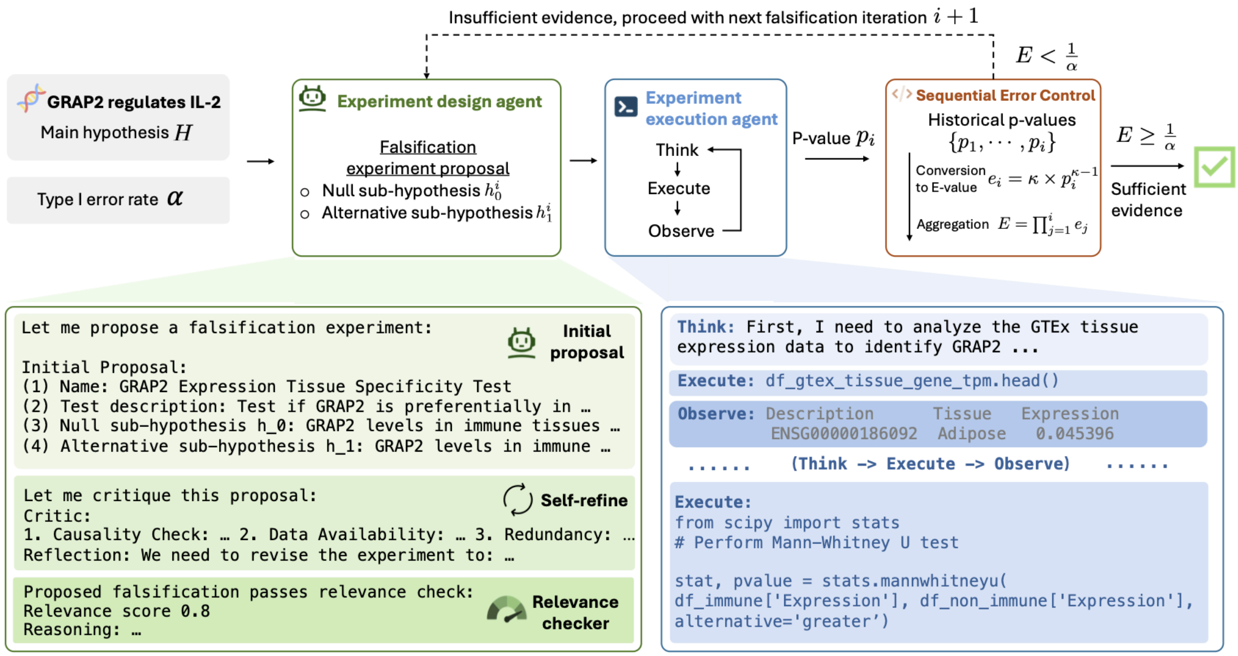
POPPER is an **agentic framework** that automates hypothesis validation using rigorous statistical principles and **LLM-based agents**. Inspired by **Karl Popper's principle of falsification**, it actively **tries to disprove** hypotheses rather than prove them.

🌍 How Does POPPER Work?

POPPER employs two specialized **AI-driven agents**:

1. **Experiment Design Agent** – Formulates falsification experiments.
2. **Experiment Execution Agent** – Conducts experiments and analyzes results.

📚 POPPER's Process



🔢 Performance in Six Scientific Domains

- **Type-I error rates below 0.10** across all datasets.
- **3.17x improvement** in validation power compared to existing methods.
- **10x faster** validation than human researchers.

💡 Case Study: Biological Hypothesis Testing

In a study on **Interleukin-2 (IL-2) and immune response**:

- POPPER's testing mechanism **outperformed Fisher's combined test**.
- Achieved **expert-level accuracy** in hypothesis validation.
- **Reduced validation time by 10-fold**, proving its efficiency.

🥇 Expert Validation

Evaluated by **9 PhD-level computational biologists & biostatisticians**, POPPER:

- Matched **human performance**.
- **Dramatically reduced** the time needed for hypothesis validation.