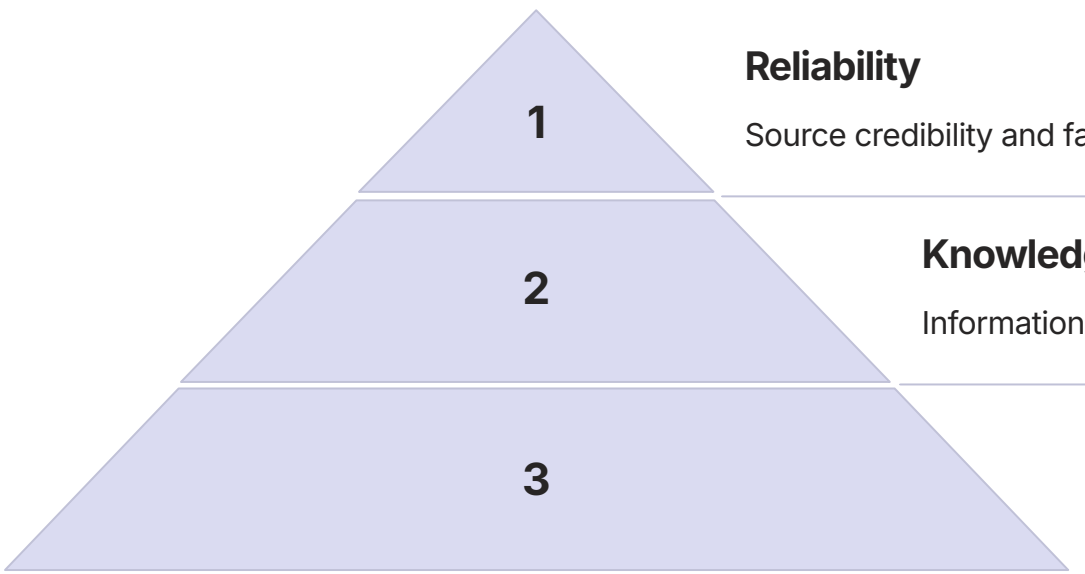
Thematic Cluster 3: Knowledge, Retrieval & Robustness

This cluster focuses on the **reliability and structure of knowledge** in LLMs and retrieval-augmented systems. Projects analyse the robustness of retrieved information, knowledge compression or distillation, and mechanisms by which models verify, maintain, or adapt factual content across domains.



Reliability

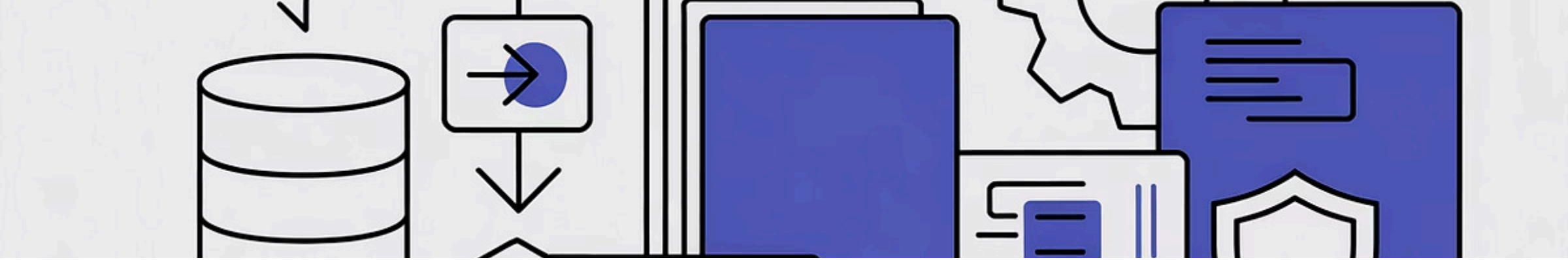
Source credibility and factual verification

Knowledge Structure

Information organization and representation

Robustness Analysis

System resilience against misinformation and conflicting sources



P5. In RAG We Trust?

This project evaluates how **Retrieval-Augmented Generation (RAG)** systems manage source credibility and factual reliability when retrieving information from multiple documents. Instead of designing new RAG architectures, students focus on **quantifying and analyzing reliability** by measuring how models weigh and reconcile conflicting or falsified sources.

Core Pipeline

RAG pipeline with intentional "poisoned" retrieval component. Test robustness against false or conflicting evidence while prompting source reliability assessment.

Expected Outcomes

Quantitative assessment of RAG system misinformation handling, producing metrics and qualitative insights on trust mechanisms.

Methodology

- **Controlled Data Poisoning**

  Introduce intentional falsehoods or contradictions into retrieved document sets

- **Multi-source Verification**

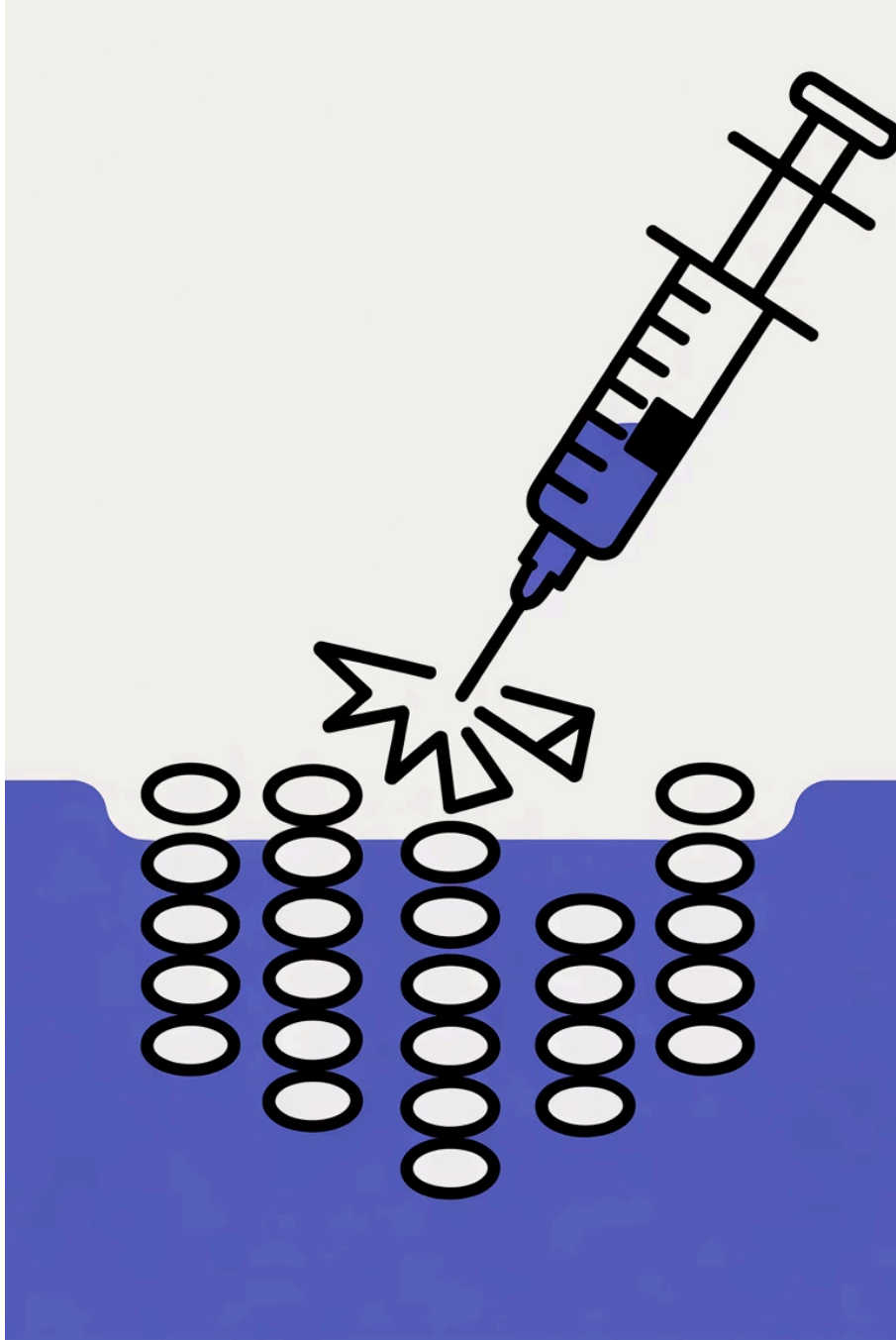
  Measure whether model identifies inconsistencies, seeks confirmation, or hedges answers

- **Prompt Design**

  Test whether meta-prompts improve reliability (e.g., "check consistency across documents")

- **Evaluation**

  Compare factual accuracy, hallucination rate, and self-consistency across models and retrieval settings



Dataset & References

1

FEVER Dataset

Fact-checking and verification benchmark for evidence-based reasoning

2

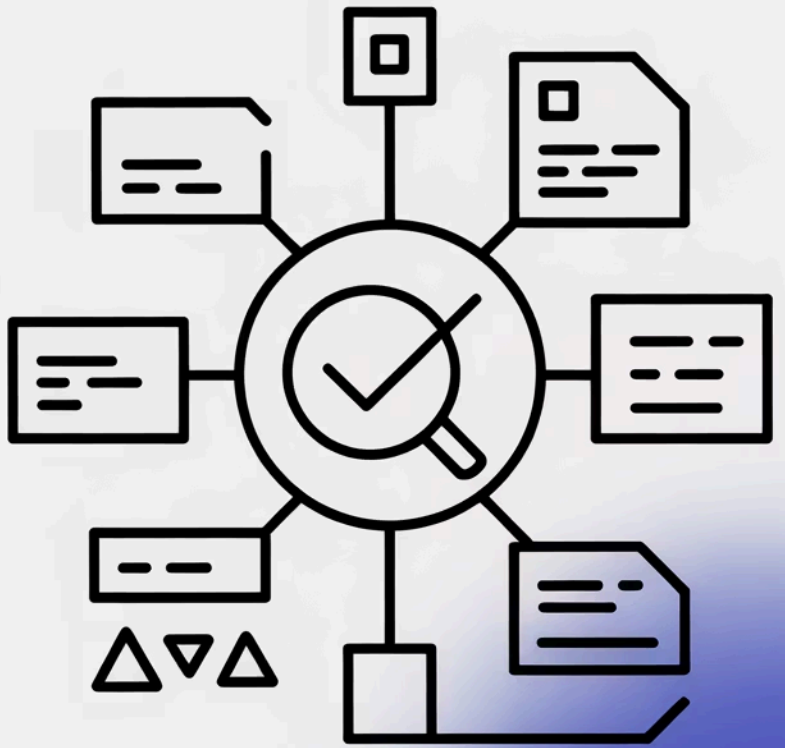
HotpotQA

Multi-hop QA benchmark for evidence reasoning and source verification

3

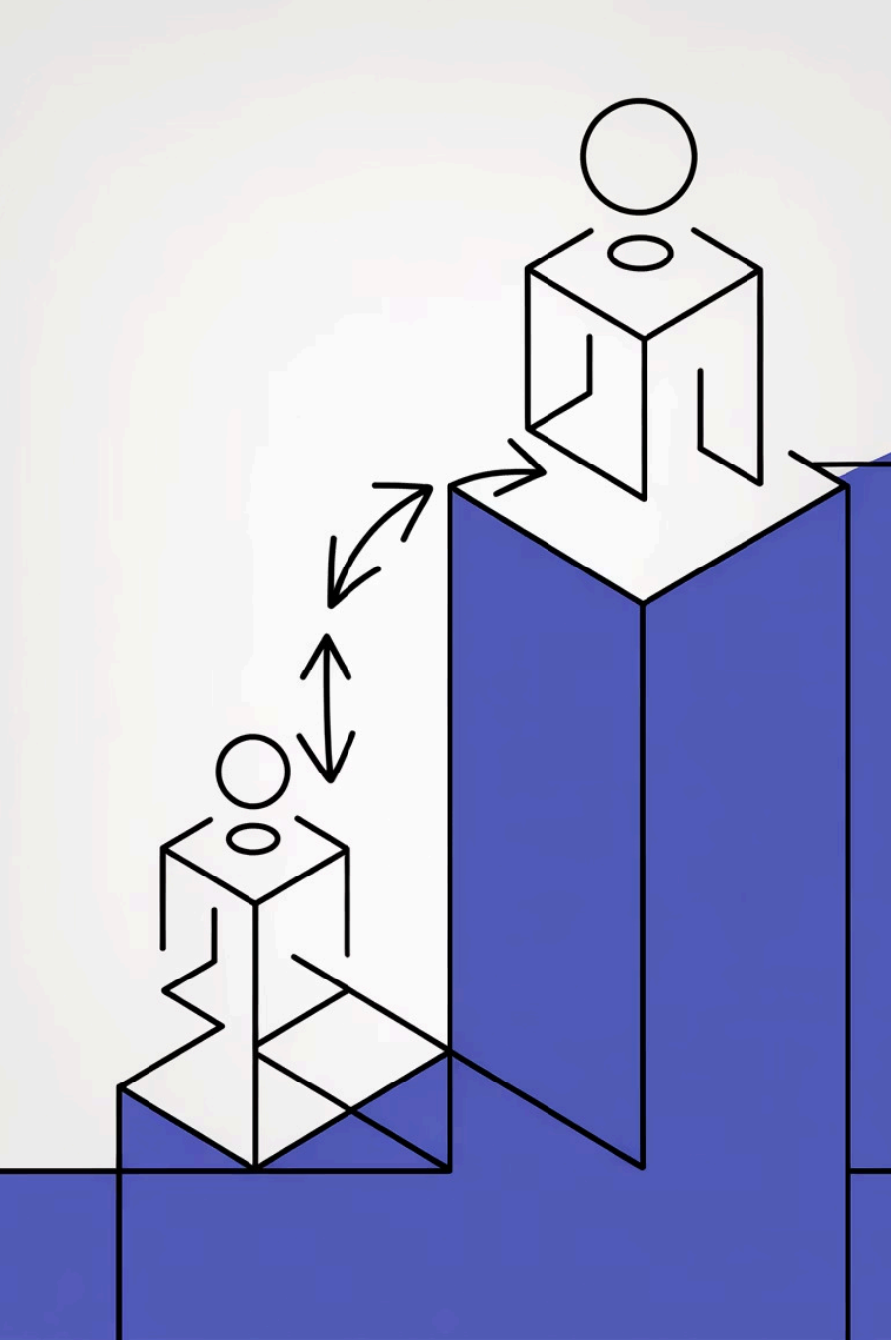
Synthetic Datasets

Custom datasets with deliberately injected false or contradictory passages



References

- Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- Zhou, Y., et al. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.
- Singal, R., et al. (2024). Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. *FEVER Workshop*, 91-98.



P6. The Apprentice Model

This project explores **knowledge distillation** by training a smaller model to imitate a larger LLM on a specific domain or reasoning task. The aim is to obtain lightweight, domain-specialized "expert" models while analyzing trade-offs between **efficiency, specialization, and generalization**.

1

Core Pipeline

Smaller model trained using outputs or intermediate representations from teacher LLM. Student predictions compared with teacher across benchmarks.

2

Expected Outcomes

Measure trade-offs between model compactness, domain specialisation, and reasoning quality retention.