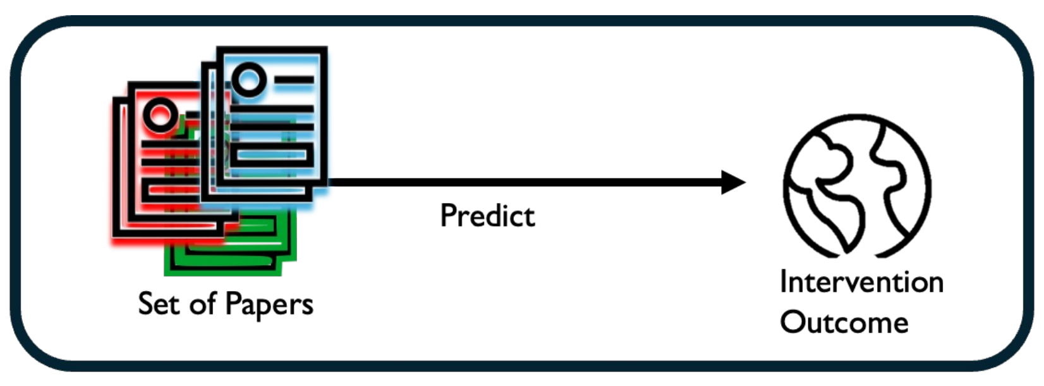
# Language Models as Tools for Research Synthesis and Evaluation

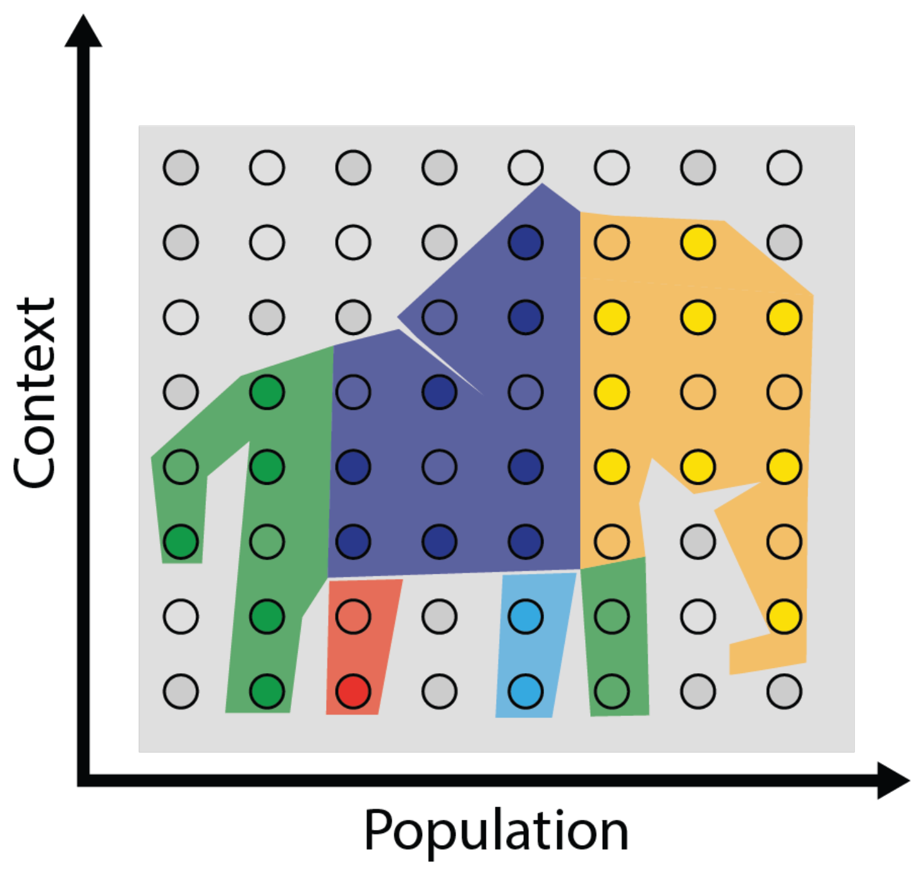
*Robin Na, *Abdullah Almaatouq
*Massachusetts Institute of Technology

## Meta-scientific Motivation

- Evaluating **the cumulative progress** in scientific findings is challenging, despite histories of developing methods to evaluate the internal validity of a scientific paper within a specific context
- Most problems in the social and behavioral sciences show high contingency and context dependency, leading to high heterogeneity in treatment effects that one research paper likely does not fully capture.



*Can we utilize LLMs and their retrieval ability of academic papers to evaluate the informativeness of each set of papers in predicting heterogeneous treatment effects?*

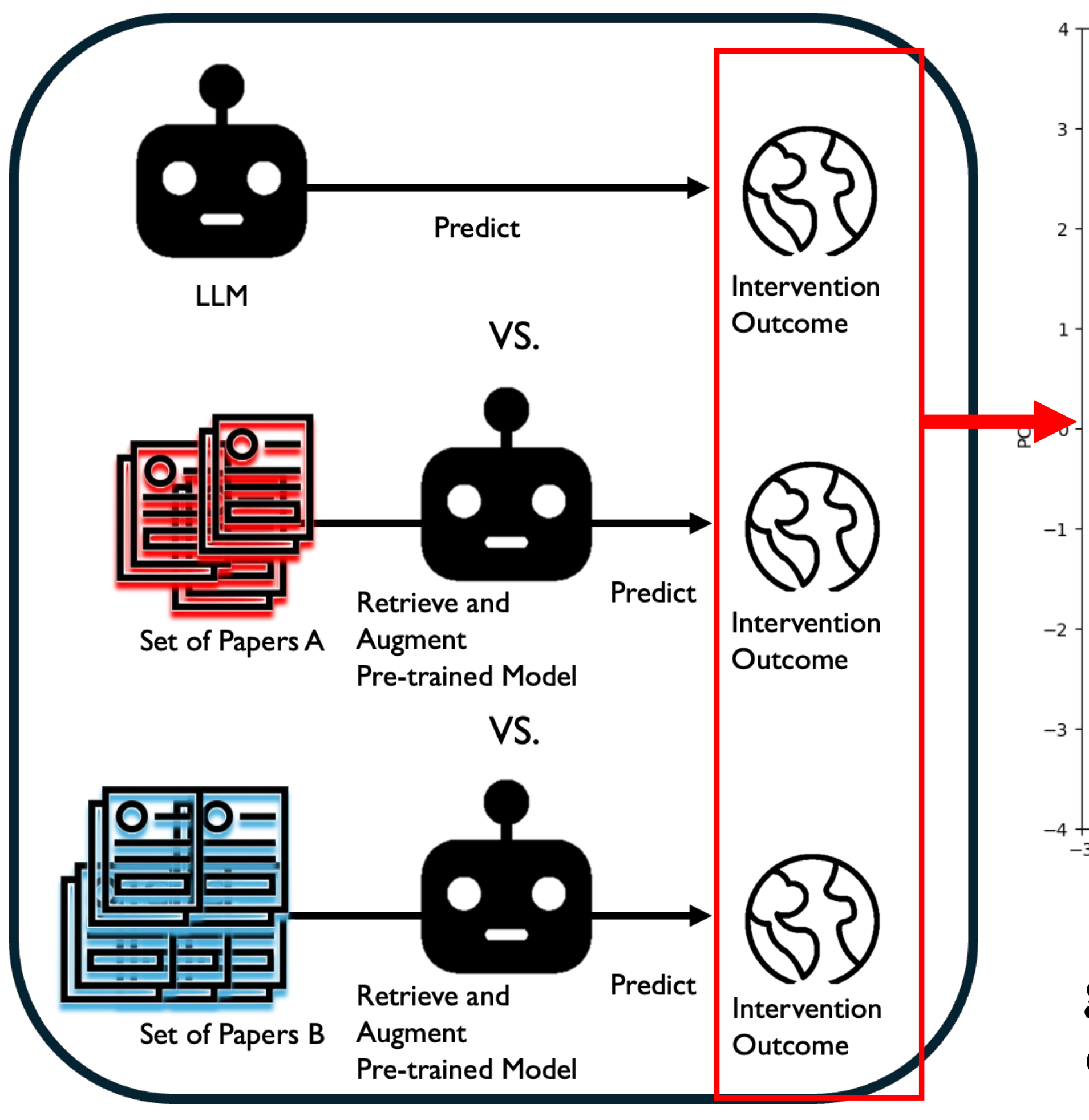


## Challenges

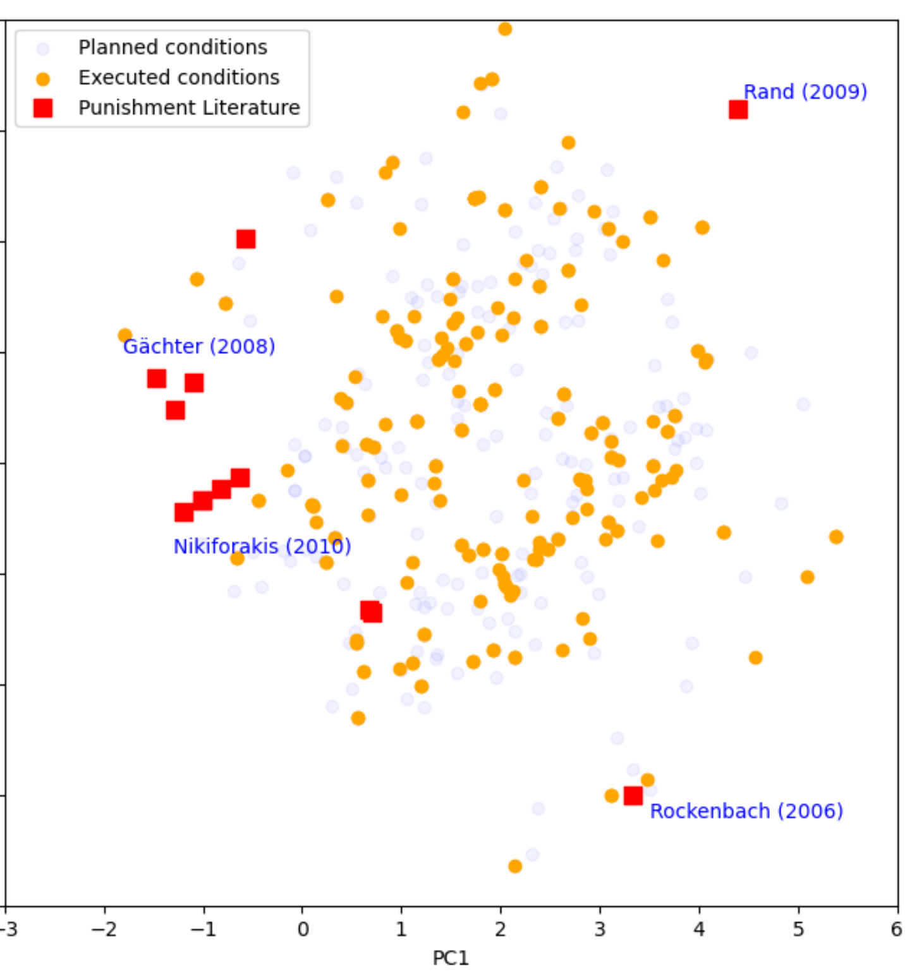
- **Synthesis:** each study is highly contextual. The assumptions and environments within a scientific finding is often implicit. How do we map each paper into the high-dimensional context space? How do we address between competing theories and findings (incommensurability problem)?
  - We need a more sophisticated meta-analysis tool that goes beyond statistically aggregating the results across papers.
- **Evaluation:** most available data from previous studies are collected from a narrow, specific context, not a systematic sample across various contexts.
  - We need a more comprehensive benchmark.

## Our Framework: LLM+RAG for literature evaluation

We develop a **framework** and a **benchmark** for addressing these challenges



Compare the quality of retrieval documents (set of academic papers) by eliciting predictions from retrieval-augmented LLMs

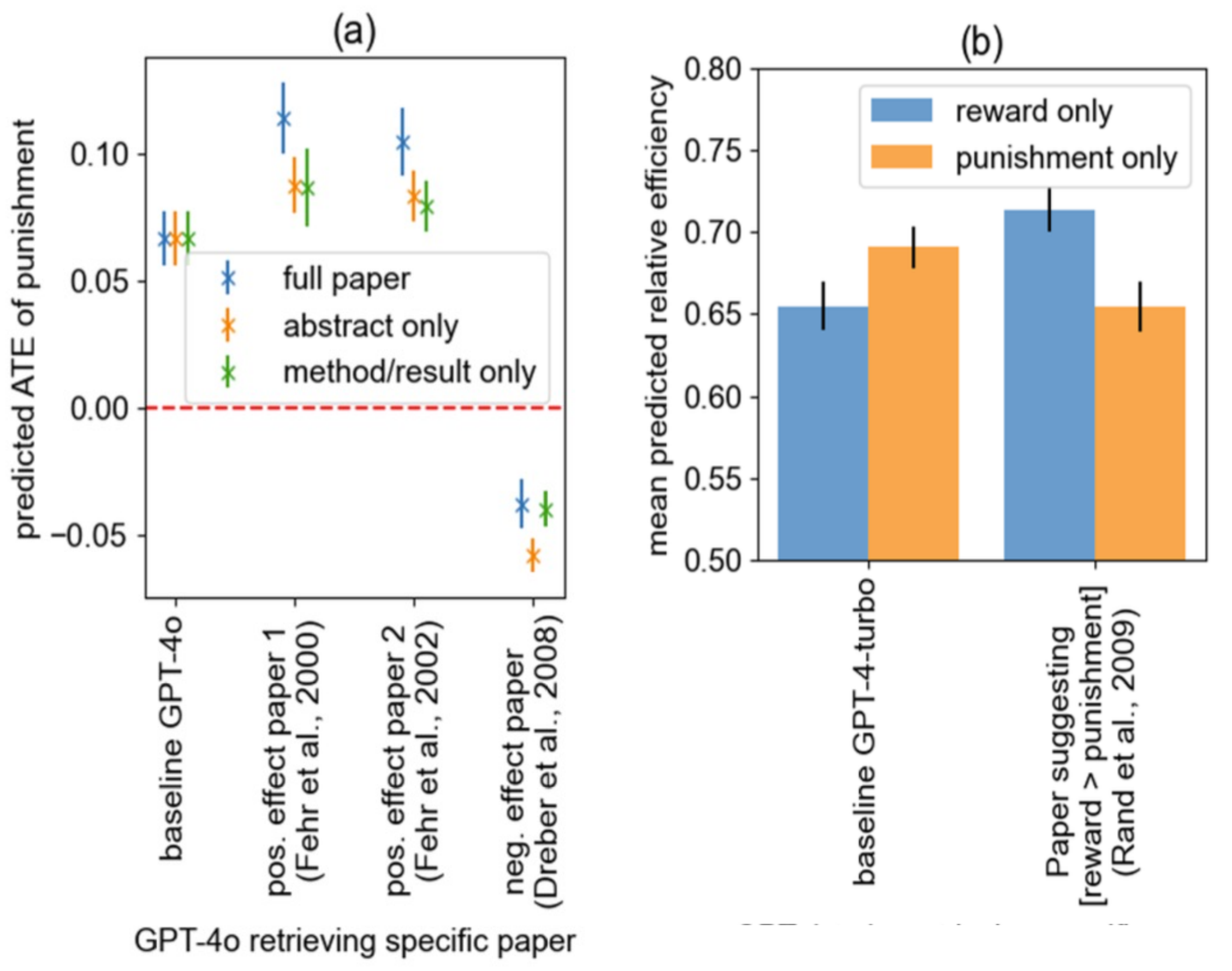


**Evaluation Set:** 211 public goods game (PGG) experiments conducted across 20 experimental configurations (e.g., number of players, information display), following integrative experiment design (IED)

**Verification steps for extracting the signal of the literature, not the language model itself or its retrieval ability**

- Step 1: Factual retrieval – is the LLM able to retrieve relevant context of the research (e.g., experimental design) and the result from each paper?
- Step 2: Ranking – is the LLM able to rank multiple articles based on each of its expected relevance to the its prediction scenario?
- Step 3: Manipulation check – does the LLM's changes in prediction align with the implication of retrieved academic papers?

## Manipulation Checks

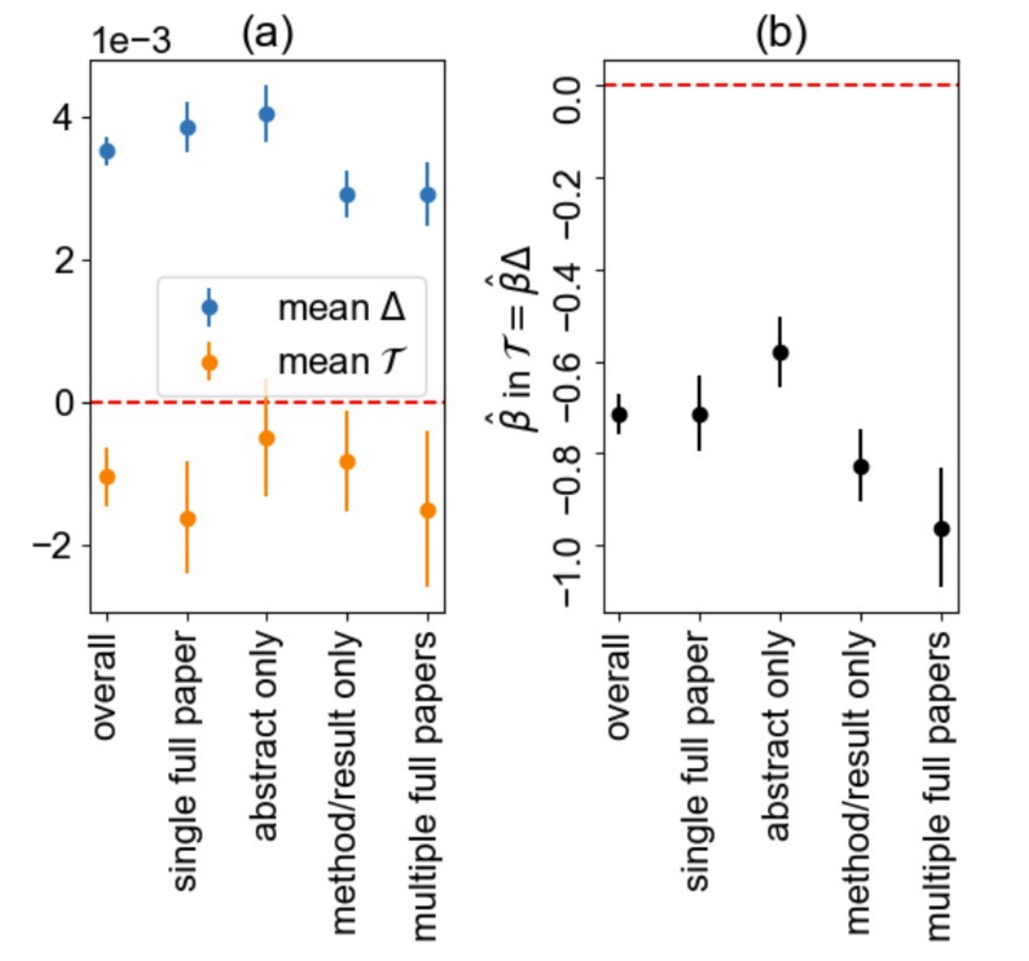


## Literature Treatment Effect: Overall

$$\mathcal{T}_i^j = \mathcal{L}(y_i^{true}(X_i), \hat{y}_i^{LLM}(X_i)) - \mathcal{L}(y_i^{true}(X_i), \hat{y}_i^{LLM}(X_i; L_j))$$

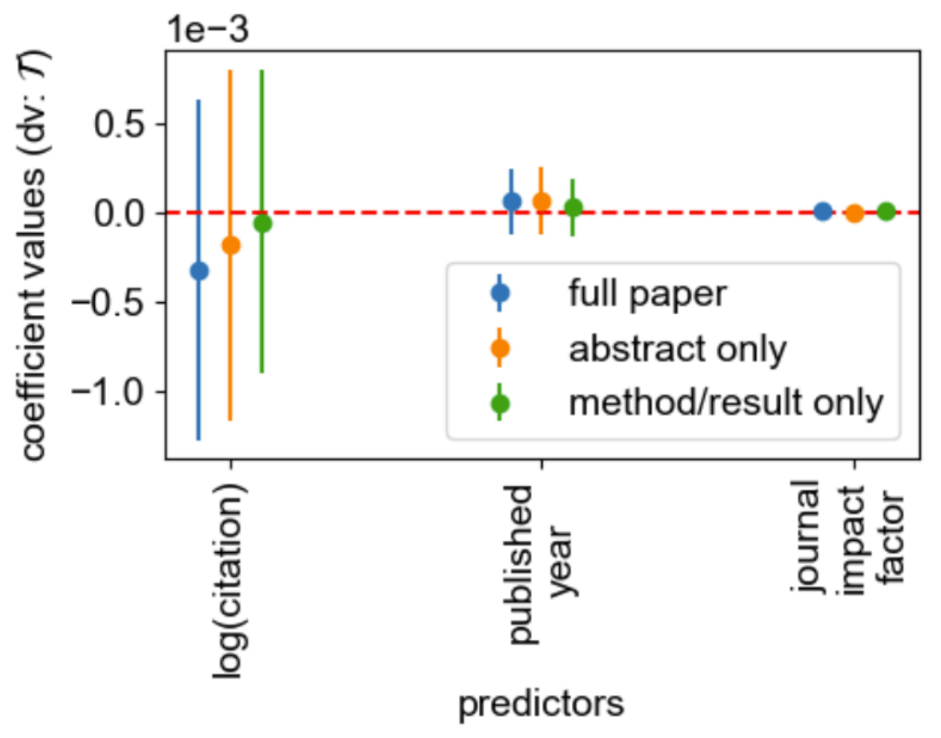
$$\Delta_i^j = \mathcal{L}(\hat{y}_i^{LLM}(X_i; L_j), \hat{y}_i^{LLM}(X_i))$$

Measuring how for each experiment $i$, reading a set of documents $L_j$ improves ($T$) and changes ($\Delta$) LLM's prediction $\hat{y}_i$ based on context vector $X_i$

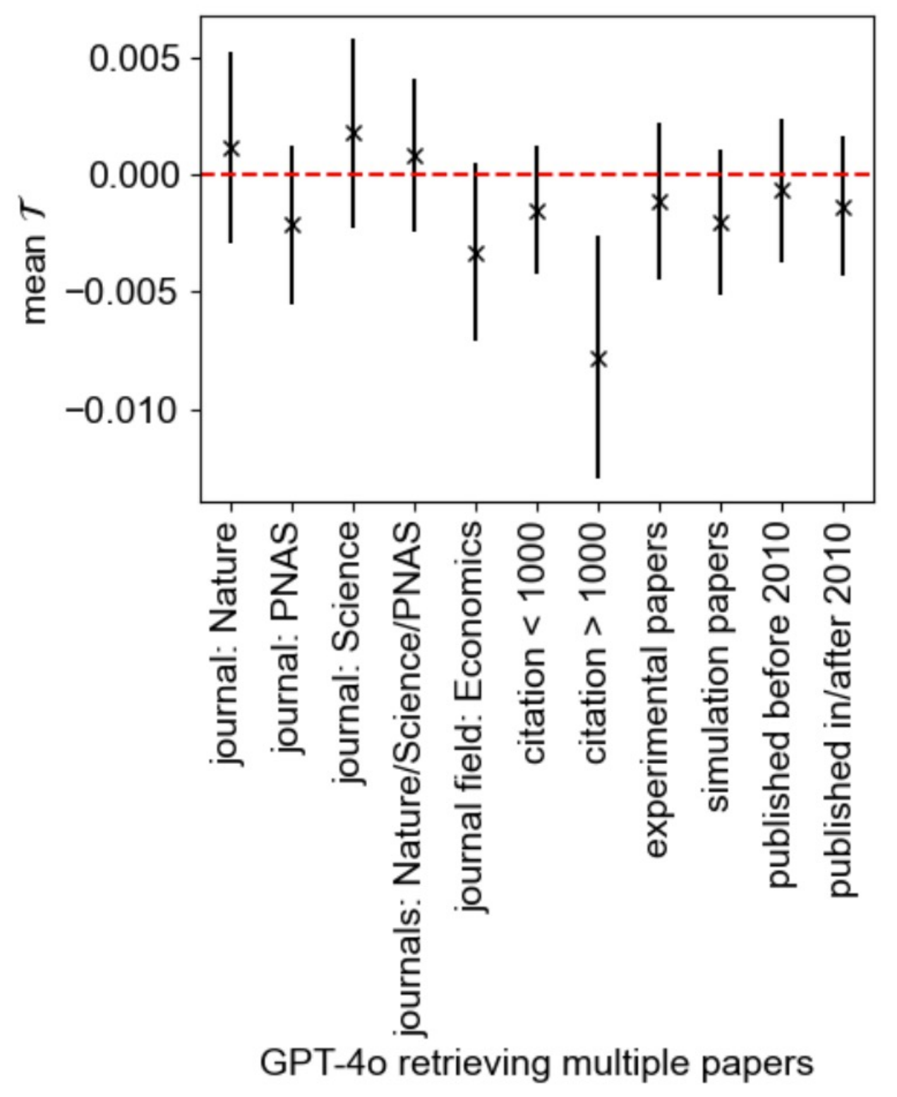


Retrieving papers generally worsens the predictive accuracy of GPT4, especially when the shift in response is drastic.

## Literature Treatment Effect: By Paper Features

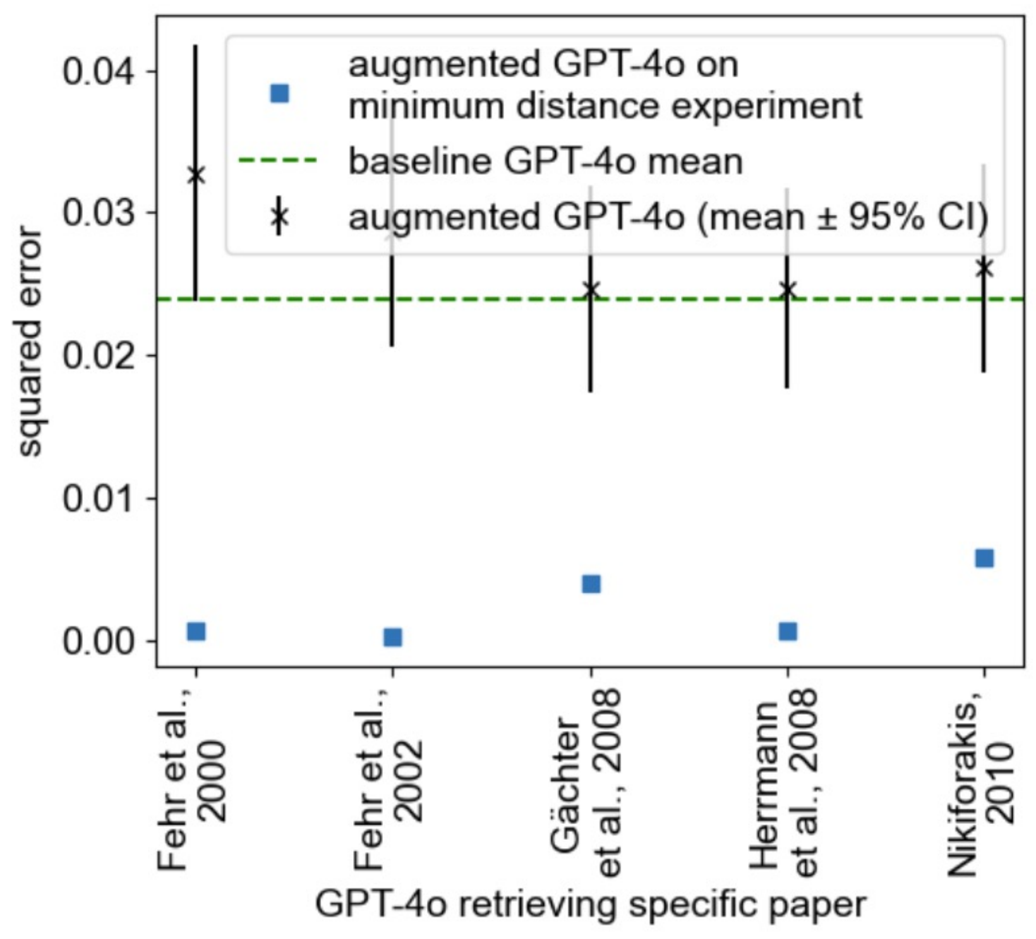


Features of individual papers do not predict the improvement.



No subset of the 25 prominent papers in the topic show significant improvement on GPT4's prediction.

## Internally Valid, Externally Invalid



The LLM performs well on tasks where the experimental designs of the test set and the retrieval sources are aligned, but not on ones that deviate.

This not only implies the high heterogeneity and sensitivity of PGG experiments depending on experimental design, but also how our framework can be useful in evaluating the paper's alignment with various unforeseen social science instances.

## Discussions

- It has been demonstrated empirically that performing RAG on unreliable documents worsen the performance of LLM. Can we flip this around and evaluate the reliability of scientific documents, going beyond the traditional scientometrics?
- Does open science movements such as preregistration and replication checks improve the predictive contribution of papers? One way to operationalize!

## Future Work (in progress)

- Comparing the result with human expert/laypeople predictions
- Robustness checks across different predictor and retriever models
- Scaling up the size and diversity of academic retrieval documents