GENEVA GRADUATE INSTITUTE

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT

GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Graduate Institute of International and Development Studies

International Economics Department

Working Paper Series

Working Paper No. HEIDWP14-2026

Following the Crowd: Literature Support and the Capabilities of Autonomous Research Agents

Michele Zampa
Geneva Graduate Institute

Chemin Eugène-Rigot 2
P.O. Box 136
CH - 1211 Geneva 21
Switzerland

Following the Crowd: Literature Support and the Capabilities of Autonomous Research Agents

Michele Zampa*

May 18, 2026

Abstract

Machine-learning models often perform poorly when asked to generalize beyond the support of the training distribution. This paper asks whether the same limitation shapes the research capabilities of autonomous large language model (LLM) agents: do they perform better when generating papers that follow research paradigms already well represented in the literature? I study this question using evidence from the Autonomous Policy Evaluation (APE) project, an open platform developed by the Social Catalyst Lab at the University of Zurich in which LLM agents generate empirical economic policy papers and compete in a tournament-style evaluation against human-written benchmarks. I construct a measure of literature support by locating each paper abstract in the semantic space of economics using a comprehensive corpus of English-language economics abstracts from OpenAlex. This measure captures whether a paper lies in a crowded or sparse region of the discipline's existing research landscape. I then test whether literature support predicts tournament performance. I find that literature support shows a statistically significant positive association with performance for AI-generated papers, but not for human-written papers. Because outcomes are assigned by an LLM judge, this relationship could partly reflect evaluation bias toward more familiar topics. However, the absence of a comparable pattern among human papers suggests that the result is not purely judge-side. The evidence is more consistent with a production-side interpretation: autonomous research agents perform better when operating in areas that are more densely represented in the existing literature and, plausibly, in model training data. The findings shed some light on both the promise and the limits of agentic LLM systems as producers of scientific research.

Keywords: LLM agents; AI-generated research; economics of science; semantic embeddings; scientific novelty

JEL Codes: O33, B41, A14, D83

*Geneva Graduate Institute. E-mail: michele.zampa@graduateinstitute.ch

1 Introduction

Given the recent rise of easily deployable agent-level frameworks for large language models (LLMs), such as Claude Code and OpenAI Codex, there has been growing interest in whether these systems can automate substantial parts of the research process in economics and the social sciences more broadly. Much of this optimism reflects the fact that agentic systems often perform better on complex tasks than single-prompt interactions with standalone models. This improvement, however, does not primarily come from a dramatic change in the underlying reasoning capabilities of base models. Rather, it comes from architectural scaffolding that enables iterative planning, task decomposition, tool use, and intermediate validation, allowing models to approach difficult problems through a more structured divide-and-conquer process.

This paper starts from a simple hypothesis grounded in the machine-learning literature. If modern models generalize poorly outside the support of the distributions on which they are trained, then autonomous LLM research agents should perform better when generating papers that follow paradigms, topics, and empirical templates that are already familiar from the existing literature. Conversely, they should perform worse when asked to produce research in semantically sparse or genuinely novel areas. The paper uses the Autonomous Policy Evaluation (APE) project to test that hypothesis. APE is an open platform developed by the Social Catalyst Lab at the University of Zurich in which LLM agents generate empirical economic policy papers and compete in a tournament-style evaluation conducted by an LLM judge.

To operationalize this idea, I construct a semantic-density indicator that measures how strongly each APE paper is supported by the existing economics literature. The measure is built from a large corpus of English-language economics abstracts from OpenAlex, an open and continuously updated database of scholarly publications, authors, institutions, and research topics. Using all abstracts classified as economics from 2000 onward, I estimate whether each APE paper abstract lies in a crowded or sparse region of the discipline's semantic space. In substantive terms, the measure captures whether a paper is located near well-populated and familiar research paradigms or instead in a thinner and less represented part of the literature. I then test whether this measure of literature support predicts tournament performance.

The main result is clear: literature support is positively associated with tournament performance for AI-generated papers, but not for human-written papers. This is the pattern one would expect if autonomous research agents are more effective when they can stay close to research paradigms that are already well represented in the underlying distribution of past text and ideas on which they were trained. Another possible interpretation is judge-side familiarity. Because tournament outcomes are evaluated by an LLM, papers located in denser and more familiar regions of the literature may appear more plausible and therefore receive higher scores. Yet the absence of a comparable relationship for human-written papers suggests that this mechanism alone cannot explain the pattern. The evidence is more consistent with a production-side interpretation, namely that LLM research agents themselves perform better when generating work

in areas that are more densely represented in existing research and, plausibly, in their training data. Still, because the APE competition includes only a small number of human-written papers, this interpretation should be treated as suggestive rather than definitive.

This contribution speaks to two related literatures. The first studies the usefulness of LLMs within social-science workflows. Current evidence suggests that LLMs are already valuable for several bounded components of research, especially structured text-processing tasks such as classification, labeling, and explanation generation, where outputs can be locally validated against human judgments or gold-standard datasets. In these settings, LLMs can match or exceed crowdworker performance and perform competitively across a wide range of computational social-science benchmarks, even if they generally do not outperform task-specific supervised models when those are available (Gilardi et al., 2023; Ziems et al., 2024; Törnberg, 2024). By contrast, the literature offers much weaker support for treating LLMs as reliable autonomous scientific reasoners. Methodological evaluations continue to raise concerns about construct validity, prompt sensitivity, dataset contamination, and instability across inference settings, all of which limit the use of these systems for theory building or causal inference without external validation frameworks (Abdurahman et al., 2025; Ludwig et al., 2024). Related work shows that LLM-based agents can assist with causal-structure proposal, simulation-based hypothesis testing, and research scaffolding when embedded in structural causal-modeling pipelines, but remain unreliable for estimating causal effect magnitudes without conditioning on fitted models and do not yet support end-to-end autonomous research pipelines in dynamic settings that require persistent reasoning and epistemic calibration (Manning et al., 2024; Kıcıman et al., 2024; Kim et al., 2025). Taken together, the strongest current evidence supports the use of LLMs as workflow components and analytical assistants rather than as autonomous producers of social-scientific knowledge.

The second literature concerns why these limitations persist, and it provides the core premise behind the paper's hypothesis. A central possibility is that the problem reflects a deeper generalization failure that has been documented across modern machine learning more broadly: models often rely on shortcuts, spurious correlations, and interpolation within the support of past data rather than robust extrapolation to genuinely novel settings (Geirhos et al., 2020; Recht et al., 2019; Koh et al., 2021; Tu et al., 2020). Recent large-scale evidence suggests that even apparent zero-shot performance in multimodal foundation models largely tracks concept frequency in pretraining corpora rather than genuine out-of-distribution reasoning ability: downstream performance improves only with exponentially increasing exposure to relevant concepts during training, and it deteriorates sharply on long-tailed or synthetic distributions designed to remove such exposure (Udandarao et al., 2024). Consistent with this interpretation, recent LLM-specific studies show that even highly capable models remain weak at simple relational reversal and at extrapolating to larger or structurally different environments unless additional task-specific training or scaffolding is introduced (Berglund et al., 2023; Kim et al., 2024), and that apparent step-by-step reasoning performance can degrade sharply once problems move outside familiar distributional structure, producing what

has been described as an "illusion of thinking" rather than stable compositional inference (Shojaee et al., 2025). Complementary evidence further suggests that probabilistic text generation itself introduces measurable structural regularities in model outputs: using lossless compression as a model-agnostic diagnostic, Hadad et al. (2026) show that LLM-generated language exhibits systematically higher compressibility than human-written text across controlled human-LLMs continuations, knowledge infrastructures, and synthetic social interaction environments, consistent with concentration of output within highly recurrent statistical patterns rather than flexible abstraction. More broadly, this interpretation aligns with recent arguments that the apparent overlap between human and machine judgments often reflects surface-level linguistic plausibility rather than shared epistemic structure, because LLMs operate primarily as stochastic pattern-completion systems rather than agents that form grounded beliefs or causal models of the world (Quattrociocchi et al., 2025). Together, these findings suggest that the flexibility of foundation models often reflects dense coverage of web-scale training distributions rather than robust abstraction or transferable scientific reasoning.

Against this background, the paper provides a direct empirical test of whether autonomous research agents perform better when they can lean on familiar research paradigms rather than operate outside the effective support of what they have seen before. Rather than evaluating isolated benchmark tasks, it studies a setting in which agentic systems generate full research papers and are assessed in competitive comparisons. The results indicate that literature support matters for AI-generated papers, but not for human-written papers. This pattern provides suggestive evidence as to where autonomous research agents are currently strongest, where their limits are likely to appear, and why claims about end-to-end research automation should be interpreted with caution.

2 Data Sources

2.1 The Autonomous Policy Evaluation project

This paper draws on research outputs generated within the APE project, an open experimental platform developed by the Social Catalyst Lab at the University of Zurich to study whether artificial intelligence systems can automate empirical policy research. The central objective of the project is to evaluate whether autonomous workflows based on LLMs can generate, replicate, and iteratively improve observational policy evaluation studies using publicly available data. The broader motivation is that scalable automation of policy evaluation could substantially accelerate the identification of effective interventions by expanding the volume and speed of empirical economic analysis.

The APE system produces complete research papers end-to-end, including data acquisition, econometric analysis, and manuscript preparation, using publicly available observational datasets such as administrative statistics, surveys, and government records. Importantly, no manual data collection is performed for

individual studies; instead, data are retrieved programmatically through public interfaces and processed automatically within a reproducible pipeline (Social Catalyst Lab, 2026).

A distinctive feature of the project is its use of a tournament-style evaluation framework in which AI-generated papers are compared against human-authored research drawn from top economics journals. All generated outputs, replication materials, intermediate versions, and failures are released publicly, allowing external auditing and methodological scrutiny. The project therefore provides a uniquely transparent setting for studying the comparative performance of human and machine-generated economic research and for evaluating the epistemic properties of automated research systems.

Another important feature of the APE project is that paper evaluation is itself conducted using an LLM, specifically Gemini 3.1 Flash Lite. Within the tournament framework, the model serves as an automated referee that compares pairs of research papers and determines their relative quality according to criteria intended to approximate editorial standards at leading economics journals. The evaluation protocol instructs the model to reward papers that pose novel research questions challenging conventional wisdom, implement credible identification strategies even when results are null, and engage transparently with methodological limitations, while penalizing weak identification, failed placebo tests or violated assumptions, and shallow empirical analysis lacking appropriate robustness checks. Final rankings are computed using the TrueSkill rating system (Herbrich et al., 2007), a Bayesian skill estimation framework originally developed by Microsoft, in which each paper is assigned a mean performance parameter (μ) representing its estimated quality and an uncertainty parameter (σ) reflecting the confidence associated with that estimate; Elo scores are additionally reported as a complementary ranking metric. Papers are ordered according to the conservative score $(\mu - 3\sigma)$, which corresponds to a lower-bound estimate of performance and ensures that high rankings reflect consistent success across multiple pairwise comparisons rather than isolated or potentially noisy match outcomes.

The present study collects both AI-generated and human-authored papers from the APE website, together with their abstracts and corresponding performance scores in the tournament evaluation (including both Elo ratings and conservative TrueSkill scores)[1]. The resulting dataset includes 981 AI-generated papers and 41 human-authored papers[2]. This sample constitutes the principal dataset used to evaluate the core research question of the paper, namely whether the degree of literature support associated with an AI-generated paper is positively correlated with its performance in the APE tournament ranking.

Across the sample, human-authored papers substantially outperform AI-generated papers on average. Mean Elo scores equal 1743 for human papers and 1263 for AI-generated papers, while mean conservative TrueSkill scores equal 27.8 and 12.3, respectively. Figure 1 illustrates the distribution of performance scores

[1]These data were collected on 13/04/2026; the APE project is continuously evolving and currently runs approximately 50 matches per day, so results will be recomputed at a later date.

[2]While 1000 APE papers had been generated at the time of data collection, I exclude papers with conservative scores equal to zero because such observations correspond to papers that had not yet participated in any tournament matches and therefore contain no information about relative performance.

separately by paper type for both ranking metrics.

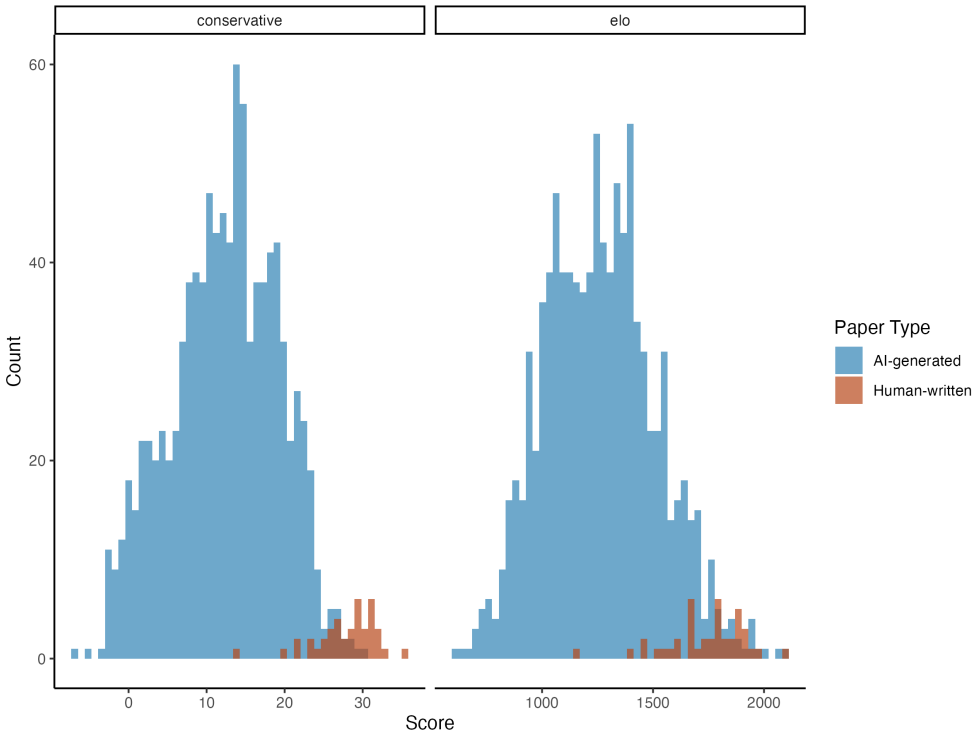

Figure 1: Distribution of APE tournament scores by paper type. The figure reports the distribution of Elo ratings and conservative TrueSkill scores for AI-generated and human-authored papers.

2.2 OpenAlex

To quantify literature support for both the generated by LLMs in the APE experiment and benchmark papers, this study relies on OpenAlex, a large-scale open bibliographic database of the global research ecosystem. OpenAlex provides structured metadata describing scholarly works, authors, institutions, venues, and concepts, as well as the citation relationships linking them.

Launched in 2022 by the non-profit organization OurResearch as a successor to the Microsoft Academic Graph, OpenAlex currently indexes hundreds of millions of scholarly outputs across disciplines and provides programmatic access through a public API and bulk data releases. The dataset represents the research system as an interconnected knowledge graph in which publications are linked to topics, references, institutional affiliations, and citation networks, enabling large-scale bibliometric analysis of scientific influence and conceptual proximity (Priem et al., 2022).

In this paper, OpenAlex is used to construct measures of literature support: the OpenAlex free API is used in order to systematically download the abstracts of all papers classified as economics between 2000 and the present day. The papers are further cleaned to retain only English-language abstracts, ensuring comparability with the APE papers, which are produced in English. The final sample contains 1.674.731 abstracts. Figure 2 reports the distribution of economics publications by year in the resulting OpenAlex sample, illustrating the steady growth in the volume of indexed abstracts over time and the strong expan-

sion of coverage after the early 2000s. This corpus forms the reference literature against which the degree of support for each APE paper is evaluated.

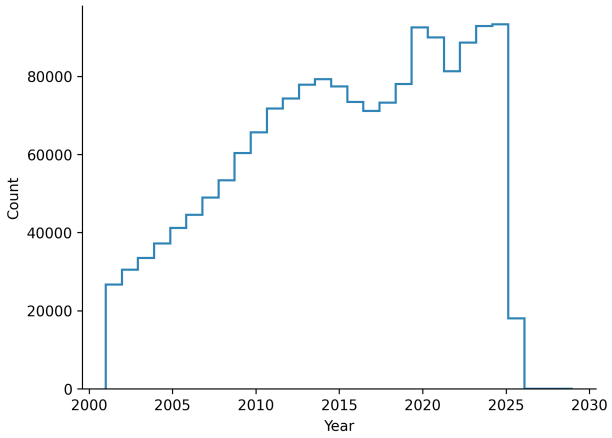

Figure 2: Distribution of English-language economics abstracts in the OpenAlex sample by publication year (2000–2026). The figure illustrates the growth in the number of indexed economics publications over time and the resulting expansion of the literature base used to construct measures of literature support.

3 Literature Support Score Based on Local Embedding Density

To measure how strongly a document aligns with existing literature, I construct a *literature support score* based on the local density of its nearest neighbors in embedding space. Intuitively, documents located in dense semantic regions receive higher support scores, while isolated or atypical documents receive lower scores.

As a first step, I construct a unified corpus of abstracts that includes both the abstracts from the APE papers and those from the OpenAlex corpus. I then embed each abstract into a vector representation using a pre-trained sentence-transformer model[3]. These embeddings map documents into a high-dimensional semantic space in which cosine similarity provides a measure of topical proximity between texts.

Let $\mathbf{e}_i \in \mathbb{R}^d$ denote the normalized embedding of document i, such that $\|\mathbf{e}_i\|_2 = 1$. For each document i, I compute cosine similarity with all other documents,

$$s_{ij} = \mathbf{e}_i^\top \mathbf{e}_j,$$

and identify the set $\mathcal{N}_k(i)$ consisting of its k[4] nearest neighbors (excluding the document itself). I then construct a kernel-weighted local density estimate over this neighborhood,

$$S_i(\tau) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \exp\left(\frac{s_{ij} - 1}{\tau}\right),$$

[3] distiluse-base-multilingual-cased-v1, available via Hugging Face.

[4] Given the size of the corpus, I set $k = 100$ to ensure computational feasibility while preserving a sufficiently local neighborhood.

where $\tau > 0$ is a bandwidth parameter controlling the sensitivity of the score to similarity differences. Smaller values of τ place greater weight on very close neighbors, while larger values produce smoother density estimates that incorporate a broader semantic neighborhood. To ensure that the results are not driven by the choice of τ, I compute literature support scores for $\tau \in \{0.05, 0.08, 0.1, 0.12, 0.15, 0.2\}$. Figure 3 illustrates the exponential similarity kernel for these alternative bandwidth values.

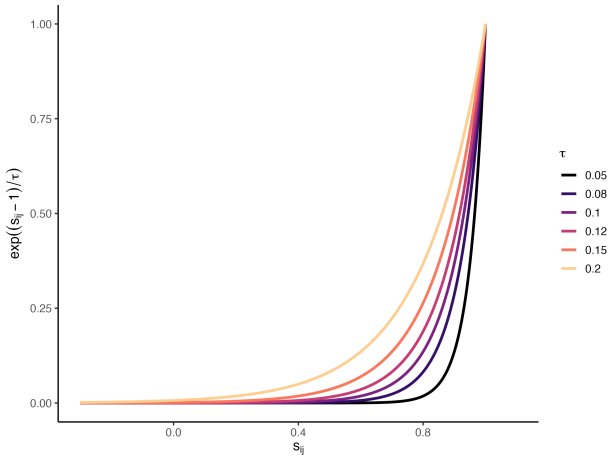

Figure 3: Kernel weighting function used to construct the literature support score. The figure plots the exponential similarity kernel $\exp((s_{ij} - 1)/\tau)$ for alternative values of the bandwidth parameter τ, where s_{ij} denotes cosine similarity between documents i and j.

Higher values of $S_i(\tau)$ indicate that a document lies in a densely populated region of the semantic space of the economics literature, whereas lower values indicate that it is relatively isolated from existing work.

For comparability across documents, I standardize the scores using a z-transformation,

$$Z_i(\tau) = \frac{S_i(\tau) - \mu_\tau}{\sigma_\tau},$$

where μ_τ and σ_τ denote the mean and standard deviation of $S_i(\tau)$ across all documents.

The resulting standardized support score $Z_i(\tau)$ measures how strongly each document is embedded within dense regions of the semantic literature space relative to the corpus as a whole. Figure 4 shows the distributions of the normalized literature support scores across different values of the bandwidth parameter τ.

4 Empirical Strategy and Results

4.1 Empirical strategy

I next test whether papers that are more strongly supported by the existing economics literature perform better in the APE tournament. Let Y_i denote the tournament outcome for paper i, measured either by the Elo score or by the conservative TrueSkill score $(\mu_i - 3\sigma_i)$. Let $Z_i(\tau)$ denote the standardized literature-support score introduced in the previous section, where larger values indicate that the paper lies in a

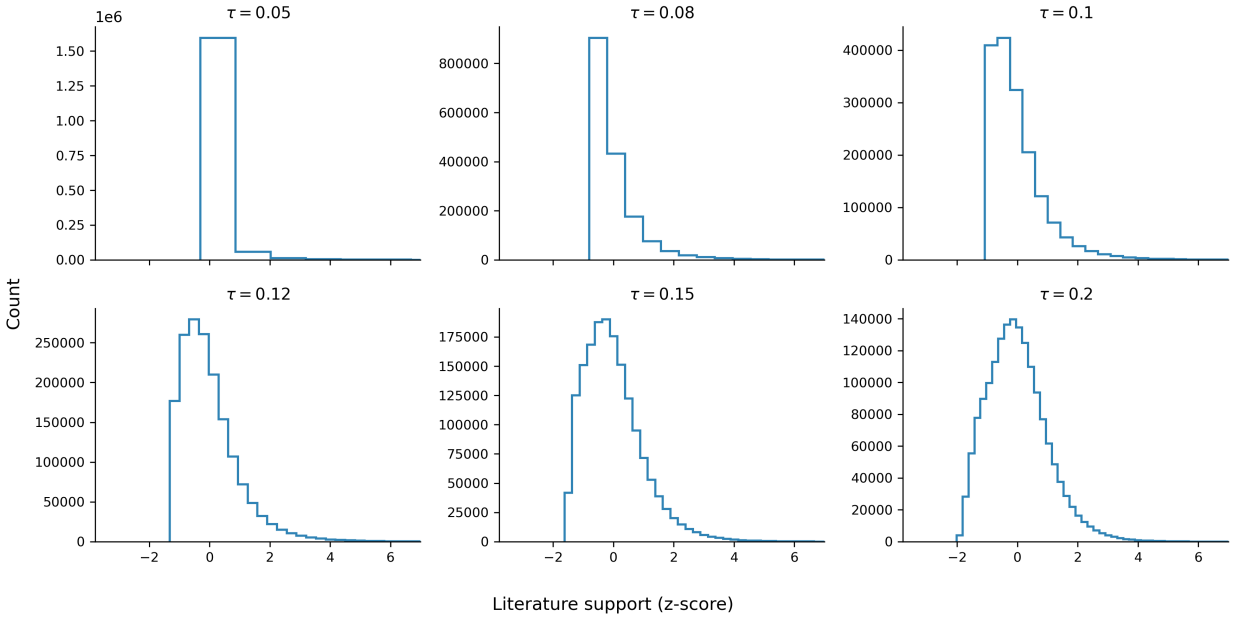

Figure 4: Distribution of the literature support score across kernel bandwidth parameters. Each panel shows the standardized distribution of the literature support measure $Z_i(\tau)$ computed using alternative values of the bandwidth parameter τ.

denser region of the semantic space of economics abstracts. I estimate the following baseline specification:

$$Y_i = \alpha_\tau + \beta_\tau Z_i(\tau) + \gamma_\tau AI_i + \delta_\tau Q_i + \lambda_\tau M_i + \varepsilon_i, \tag{1}$$

where AI_i is an indicator for whether the paper was generated by the APE system, Q_i is the share of the paper's tournament matches played against AI-generated opponents, and M_i is the total number of tournament matches observed for that paper. The inclusion of Q_i and M_i absorbs two mechanical sources of variation in scores: differences in exposure to the tournament and differences in the composition of opponents.

The baseline model identifies the average association between literature support and tournament performance across the full sample. To test whether this relationship differs between AI-generated and human-written papers, I then estimate an interaction specification:

$$Y_i = \alpha_\tau + \beta_\tau Z_i(\tau) + \rho_\tau \left[AI_i \times Z_i(\tau)\right] + \gamma_\tau AI_i + \delta_\tau Q_i + \lambda_\tau M_i + \varepsilon_i. \tag{2}$$

In this formulation, β_τ is the support-performance gradient for human-written benchmark papers, while $\beta_\tau + \rho_\tau$ is the corresponding gradient for AI-generated papers. Finally, I estimate the controlled specification separately for AI-generated and human-written papers. These split-sample regressions provide the most direct test of whether literature support predicts performance within each group.

All regressions are estimated separately for the six kernel bandwidths used to construct the support score, $\tau \in \{0.05, 0.08, 0.10, 0.12, 0.15, 0.20\}$. Because the main quantity of interest is the standardized support measure, coefficients can be interpreted as the change in tournament performance associated with a

one-standard-deviation increase in literature support. The descriptive sample contains 1,022 papers with non-zero tournament scores, but 53 of these have no recorded matches in the scraped match histories and therefore cannot be assigned a value of M_i. The estimating sample for the regression analysis is therefore 969 papers, of which 928 are AI-generated and 41 are human-written.

4.2 Main results

The results point to a clear asymmetry between AI-generated and human-written papers. In the pooled specification with controls, literature support is positively associated with tournament performance, especially for Elo. At the benchmark bandwidth $\tau = 0.10$, a one-standard-deviation increase in $Z_i(\tau)$ is associated with a 35.8 point increase in Elo (95% CI: $[0.2,\ 71.3]$). For the conservative score, the analogous estimate is 0.78 points (95% CI: $[-0.22,\ 1.78]$), which is positive but not precisely estimated.

These pooled estimates, however, mask substantial heterogeneity by paper type. Once I allow the support slope to differ between AI-generated and human-written papers, the interaction term is positive throughout. At $\tau = 0.10$, the interaction estimate is 71.5 Elo points (95% CI: $[-35.0,\ 178.1]$) and 3.26 points for the conservative score (95% CI: $[0.27,\ 6.25]$). The conservative-score interaction is statistically distinguishable from zero, indicating that the positive relationship between literature support and tournament success is significantly stronger for AI-generated papers than for human-written papers.

The split-sample regressions make this pattern more transparent. For AI-generated papers only, a one-standard-deviation increase in literature support at $\tau = 0.10$ predicts a 45.0 point increase in Elo (95% CI: $[6.9,\ 83.1]$) and a 1.20 point increase in the conservative score (95% CI: $[0.15,\ 2.26]$). By contrast, for human-written papers the corresponding estimates are small and imprecise: 10.8 Elo points (95% CI: $[-34.3,\ 56.0]$) and 0.27 conservative-score points (95% CI: $[-0.74,\ 1.28]$). Figure 5 shows these separate-sample estimates for both outcomes.

4.3 Robustness and interpretation

The core result is not sensitive to the choice of kernel bandwidth. Across all six values of τ, the coefficient on literature support remains positive in the AI-only regressions for both outcomes. For Elo, the AI-only estimate ranges from roughly 32.8 to 95.2 points per standard deviation of support; for the conservative score, it ranges from about 0.84 to 2.82 points. By contrast, the human-only coefficients remain close to zero across all bandwidth choices and are never statistically distinguishable from zero. Figure 6 summarizes this stability across specifications.

Taken together, the estimates support the paper's central claim: agentic LLM systems perform better when producing research in areas that are more densely represented in the existing economics literature. The strongest evidence comes from the AI-only regressions, where the support gradient is positive for both outcome measures and stable across bandwidth choices. The pooled models are also consistent with this

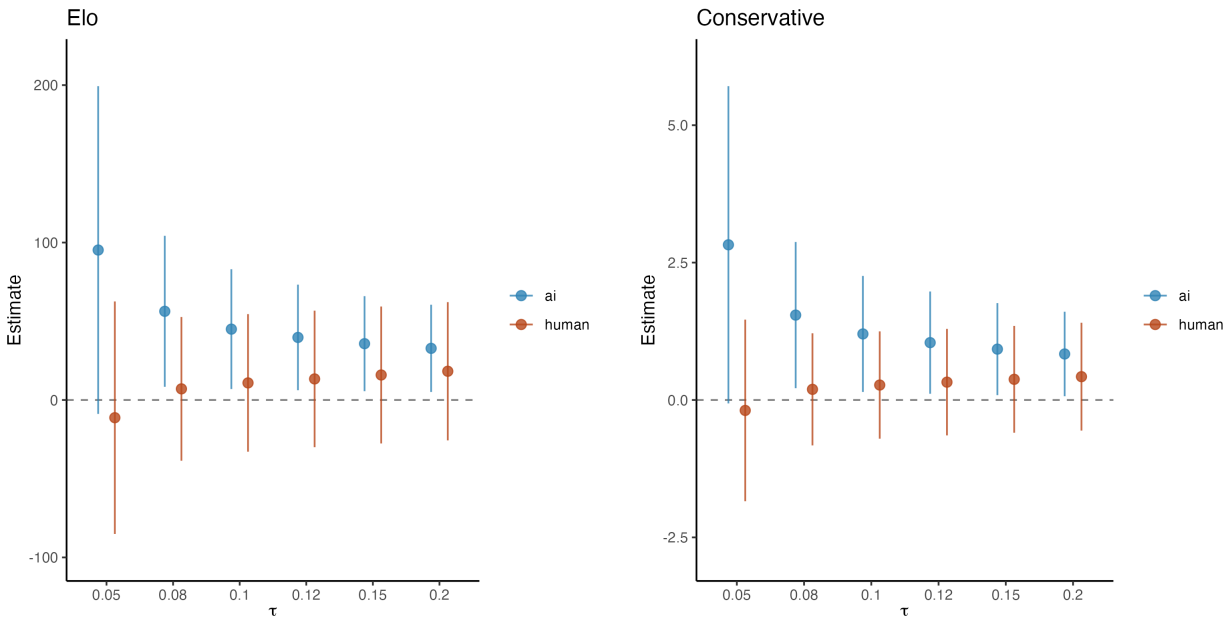

Figure 5: Estimated effect of literature support on tournament performance in separate samples of AI-generated and human-written papers. Each point reports the coefficient on the standardized literature-support score, and bars show 95% confidence intervals. The left panel uses Elo as the outcome; the right panel uses the conservative TrueSkill score. Across bandwidth choices, the support slope is consistently positive for AI-generated papers and close to zero for human-written papers.

interpretation, but they are less informative because they average over AI-generated and human-written papers. The interaction results further suggest that the support-performance relationship is stronger for AI-generated papers, especially when performance is measured using the conservative TrueSkill score.

4.4 Limitations and Measurement Caveats

Several limitations of the design should qualify the interpretation of the results. First, the outcome variable is based on judgments of complete papers, whereas the literature-support measure is constructed from abstracts only. Abstracts plausibly capture the main research question, framing, and topical positioning of a paper, and are therefore informative about whether a study sits in a crowded or sparse region of the literature. At the same time, they do not fully reflect many features that likely matter for tournament performance, including the credibility of the empirical design, the quality of the robustness analysis, the treatment of limitations, and the clarity of the full manuscript. The support measure should therefore be interpreted as a noisy proxy for the semantic location of the paper rather than as a complete representation of everything the judge observes.

Second, the literature-support score is itself only an indirect proxy for the quantity of relevant prior material available to the language model. In substantive terms, the object of interest is not simply whether a topic is common in economics, but whether it is well represented in the data environment from which LLMs learn reusable patterns, templates, and conceptual associations. The OpenAlex-based measure is useful for this purpose because it approximates the density of publicly available economics research in

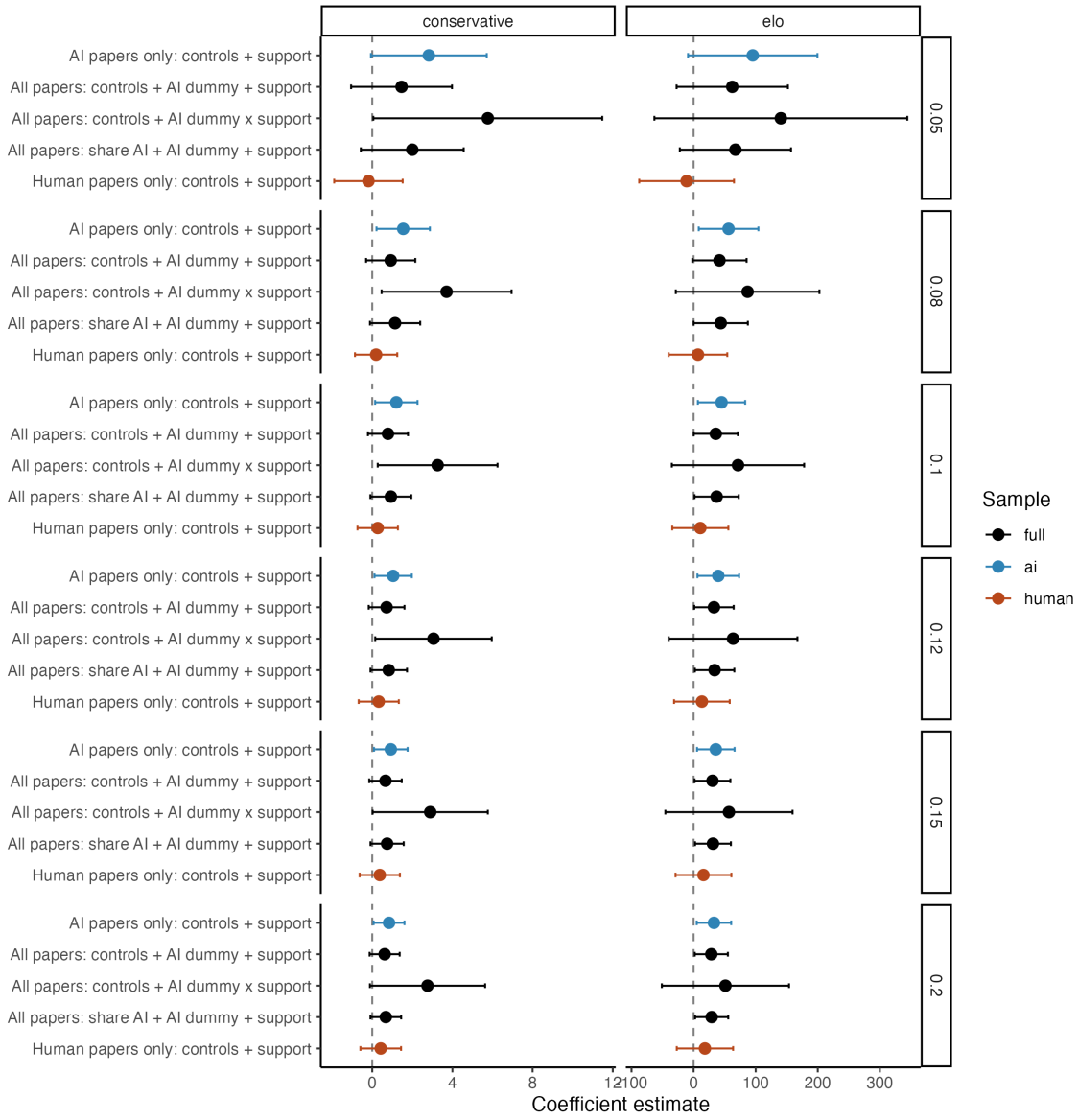

Figure 6: Literature-support coefficients across specifications, outcomes, samples, and kernel bandwidths. Black points report pooled regressions, blue points report AI-only estimates, and orange points report human-only estimates. For the interaction specification, the plotted coefficient is the differential slope for AI-generated papers relative to human-written papers.

semantic space using a large and systematically collected corpus of English-language abstracts. But it is not a direct observation of model training data, nor can it recover how a given model weights different sources, memorizes recurring structures, or combines textual fragments during generation. Given the black-box and stochastic nature of current LLM systems, the measure is best understood as a noisy indicator of latent training-distribution support rather than as a direct measure of the model's internal knowledge.

Third, the evidence remains observational and the institutional setting complicates causal interpretation. The APE experiment is still ongoing, so the paper should be read as an early snapshot rather than a final assessment of the platform. More importantly, since LLMs appear on both sides of the evaluation process, the observed association between literature support and performance could arise because AI systems genuinely produce stronger papers when working in well-supported regions of the literature, because

the LLM judge is more favorable toward papers that resemble familiar areas of economics, or because both mechanisms operate simultaneously. The absence of a comparable support gradient for human-written papers is suggestive evidence against a purely judge-side explanation, but that comparison remains limited by the small benchmark sample.

A stronger design would therefore be needed to assess the source of the observed association more precisely, and to document how the training distribution affects the research-production capabilities of agentic LLM systems. One improvement would be to use a larger and more comparable pool of human papers, rather than relying only on recent peer-reviewed articles from top economics journals. Although APE submissions are prompted to resemble top-field-journal papers, they are generated by LLMs trained on a much broader and noisier distribution of economic writing. The relevant comparison may therefore be with a wider population of human submissions, including lower-quality, unpublished, or rejected work. Another improvement would be to use human or otherwise non-LLM evaluation procedures as a validation exercise. This would be useful partly because it is difficult to ensure that an LLM judge had not been trained on the human benchmark papers it is asked to evaluate. Even recent or not-yet-published benchmark papers are not necessarily insulated from this problem, since earlier working-paper versions may have circulated online. More importantly, non-LLM evaluation would help isolate whether denser literature support improves the performance of agentic LLMs primarily on the production side, by enabling them to generate stronger papers, or on the evaluation side, by making their outputs more legible or appealing to LLM judges. A further extension would be to construct the support measure from the semantic content of full papers rather than from abstracts alone. This would provide a richer measure of the relevant training distribution, although it would require a much larger data collection and processing effort.

5 Conclusion and Discussion

The contribution of this paper is primarily practical and positive rather than normative. The analysis provides a suggestive test of where autonomous research agents currently appear to work better, not an argument that the automation of research is necessarily desirable. Questions about whether knowledge production should be increasingly outsourced to proprietary AI systems involve institutional and political considerations that lie outside the scope of the present empirical design. Those considerations include unequal access to frontier models across researchers and countries, dependence on privately controlled infrastructures, and the possibility that public research resources are increasingly redirected toward commercial platforms.

That said, if the pattern documented here generalizes, it has important implications for how research automation is likely to reshape scientific work. The results suggest that LLM-based agents may be most effective at producing work that remains close to well-established literatures, familiar empirical templates, and densely represented semantic domains. In that case, these systems may substantially expand the

throughput of incremental and standardized research while remaining much weaker at generating genuinely novel questions, shifting conceptual frames, or producing reliable work in thinly documented domains. In other words, automation may scale the production of research within existing boundaries more easily than it scales the production of new intellectual directions.

This possibility matters because the long-run effects of automation depend not only on how much research can be produced, but also on what kinds of research continue to be incentivized. If institutions respond to short-run productivity gains by reducing demand for human researchers, especially at the junior stages where many careers begin, the result could be a narrower pipeline of people developing the skills and independence required for more original work. Even if AI systems raise output in the short run, the longer-run effect could be a reduction in exploratory capacity if human researchers are precisely the margin along which more disruptive forms of innovation emerge. The broader concern, then, is not simply substitution, but selective substitution: a world in which machines become better at reproducing and extending established lines of inquiry, while the human base that generates departures from those lines becomes thinner. The results in this paper do not establish that outcome, but they do suggest that it is a possibility worth taking seriously when evaluating the role of LLMs in research and higher education.

References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., and Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8(2).

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. (2023). The reversal curse: Llms trained on "a is b" fail to learn "b is a". arXiv preprint arXiv:2309.12288.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Hadad, O., Loru, E., Nudo, J., Marco, N. D., Cinelli, M., and Quattrociocchi, W. (2026). The statistical signature of llms. arXiv preprint arXiv:2602.18152.

Herbrich, R., Minka, T., and Graepel, T. (2007). Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press.

Kim, D., Lee, J., Park, J., and Seo, M. (2024). How language models extrapolate outside the training data: A case study in textualized gridworld. arXiv preprint arXiv:2406.15275.

Kim, J., Podlasek, A., Shidara, K., Liu, F., Alaa, A., and Bernardo, D. (2025). Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports*, 15(1).

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. (2024). Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). Large language models: An applied econometric framework. arXiv preprint arXiv:2412.07031.

Manning, B., Zhu, K., and Horton, J. (2024). Automated social science: Language models as scientist and subjects. NBER Working Paper 32381, National Bureau of Economic Research.

Priem, J., Piwowar, H., and Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Quattrociocchi, W., Capraro, V., and Perc, M. (2025). Epistemological fault lines between human and artificial intelligence. arXiv preprint arXiv:2512.19466.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? arXiv preprint arXiv:1902.10811.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2506.06941.

Social Catalyst Lab (2026). Autonomous policy evaluation (ape). https://ape.socialcatalystlab.org/. Accessed: 2026-04-07.

Törnberg, P. (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Tu, L., Lalwani, G., Gella, S., and He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P. H. S., Bibi, A., Albanie, S., and Bethge, M. (2024). No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. arXiv preprint arXiv:2404.04125.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.