Research in the Age of AI: Idea Generation and Assessment in Real Estate Finance and Economics

Thies Lindenthal

Department of Land Economy, University of Cambridge

htl24@cam.ac.uk

Draft — work in progress

Last revised: 2026-04-22

AI acknowledgement:AI was used extensively throughout this project — I wanted to learn how to work with several AI services at different stages of the research process: Literature discovery (not review), data collection, setup of empirical pipelines, and econometric testing were all handled or strongly assisted by AI, mostly via Claude Code. Writing was facilitated with both Claude Code and ChatGPT. What I did myself: I developed the research idea and the empirical strategy to pursue it. I reviewed the literature, selected the papers and directed the AI agents on how to reference them in the analysis ("use this paper as an example of X"; "this paper belongs in the general literature review to make claim Y"). After auditing intermediate results I repeatedly redesigned parts of the workflow. I did not write a single line of code — which feels strange for a former programmer. I checked the empirical setup and tests. I edited the final document more extensively than I thought I would need to. The history of my instructions and prompts are available with all code at this project's repository. All errors are mine.

Abstract

Can AI act as principal investigator — formulating promising research questions — rather than just executing tasks? This paper maps the full published corpus of real estate finance and economics into a semantic embedding space, clusters it into research strands, and uses that structure to evaluate hundreds of AI-generated research ideas across eight generation conditions. Ideas are assessed using atypicality, a measure of cross-community bridging that predicts subsequent citations in retroactive validation. Econ/finance method-transfer produces the most atypical proposals; psychology-sourced conditions trail. A citation prediction model shows econ/finance conditions also leading on predicted impact, but that advantage is driven by proximity to existing work, not novelty — the two criteria are independent. Methodological scaffolding improves AI ideation; topical scaffolding does not.

---

1 Introduction

AI use in academic research is like microplastics: it is almost everywhere, and exposure is nearly impossible to avoid. Even if you do not use it yourself, it seeps in through other papers, data, and the behaviour of the researchers being studied, and you encounter it when reading, citing, and engaging with existing work. Kobak et al. (2025) applied excess-word analysis to 14 million biomedical abstracts, tracking the frequency of words that language models favour but careful human writers rarely use: "delve," "pivotal," "meticulous," "nuanced," "underscore." Rates were stable for a decade, then jumped sharply after November 2022. Their estimate: at least 13.5 per cent of 2024 biomedical abstracts were processed with an LLM, reaching 40 per cent in some subcorpora. The same analysis is replicated for roughly 101,000 papers indexed by OpenAlex under real-estate and housing concepts. The same inflection appears: "delve" runs at 0.006 occurrences per 1,000 words in 2010–2022, then jumps to 0.049 in 2023 and 0.127 in 2024 — a twenty-fold increase. The field is doing what every other field is doing. Section 5.1 documents the pattern in detail.

That is the writing layer. The more consequential shift is in the research itself. A growing subset of publications in real estate economics and finance depends on AI not for drafting but for execution — studies that could not have been done, and results that could not have been reached, without machine learning or large language models carrying out essential analytical work. Lindenthal and Johnson (2021) classify architectural styles from property photographs and link them to prices: the classifier is not a convenience but the entire study. Bartik, Gupta and Milo (2025) read and code thousands of municipal zoning ordinances, building regulation measures at national scale that no research team could produce by hand. Shen and Ross (2021) extract a description-quality measure from MLS listing text and show it carries price signals that structured data alone cannot capture. Leow and Lindenthal (2025) apply machine learning to REIT factor-return data and show substantial forecast improvements over OLS — the same approach as Gu, Kelly and Xiu (2020) in stocks, adapted to a real estate setting. In each case, AI enables a measurement or prediction that the research requires. Remove it and the paper does not exist.

These papers are counted in the corpus of *Real Estate Economics*, tracked when they appeared, and mapped to where they sit in the field's research space. The count is small but the trend is sharp: no confirmed AI/ML papers before 2020, then four in 2023 alone, reaching 4–7 per cent of annual output by 2024–2026. This estimate is a lower bound and captures only papers where AI performs a task that was otherwise infeasible — classifying images at scale, parsing thousands of zoning documents, fitting high-dimensional factor models. A larger and less visible layer of AI use runs through conventional research: data

collection and screening, literature searches, transcription, survey coding. These tasks leave no identifiable signature in the methods section and do not show up in the count, but they are almost certainly more prevalent than the visible tip. The role is that of a skilled research assistant: executing tasks specified by the researcher. The PI — the person deciding what to study and why — remains human.

This raises the question the paper pursues: how far up the value chain can AI go? An RA executes tasks specified by someone else. A PI decides what is worth doing. The difference is a specific cognitive operation: identifying questions that are novel enough to warrant investigation and feasible enough to actually pursue. Si, Hashimoto and Yang (2025a) provide the most systematic evidence to date, finding that AI-generated research ideas in NLP score higher on judged novelty than human-generated ideas but lower on feasibility, and that generation systems produce ideas that are less diverse than they appear. A follow-up study (Si, Hashimoto and Yang, 2025b) tested the execution step: when NLP researchers actually implemented both AI- and human-suggested projects, the initial novelty advantage reversed — human-suggested ideas were rated better once the work was done.

In the natural sciences, the question has moved from evaluation to execution. AlphaFold2 (Jumper et al. 2021) predicted structures for roughly 200 million proteins, a problem that had resisted structural biology for fifty years. Merchant et al. (2023) used an AI system to discover 2.2 million stable crystal structures, 736 of which were subsequently synthesised and experimentally confirmed. Boiko et al. (2023) built an autonomous agent that searched the literature, wrote code, and ran wet-lab experiments without human intervention. An AI-designed drug candidate completed a Phase II clinical trial in under four years (Insilico Medicine, 2025). In each case, the feedback loop between idea and test is directly observable: the compound either works in the animal model or it does not.

In the social sciences, no equivalent bench test exists. A research idea in real estate economics cannot be run in a laboratory, and waiting years for a paper to be written, submitted, refereed, published, and cited before judging a proposal is not an option. This asymmetry motivates the approach taken here. Rather than waiting for execution, the structure of the field's published literature — a coordinate system built from the corpus itself — serves as an ex-ante proxy for scientific value. The intuition comes from Uzzi, Mukherjee, Stringer and Jones (2013), who showed that the most-cited papers in science tend to combine a conventional base with a small number of atypical knowledge combinations.

The coordinate system is built from the published archives of *Real Estate Economics*, the *Journal of Real Estate Finance and Economics*, and real-estate-relevant subsets of the

Journal of Urban Economics and three leading general economics and finance journals, using a large language model to extract from each paper its research question, empirical method, data source, sector, and main contribution. These structured representations are embedded in a shared semantic space and clustered into research strands. The result is measurement infrastructure, not a conventional literature review — a map against which new proposals can be located relative to everything the field has already produced. The field has repeatedly advanced by importing ideas from neighbouring disciplines: hedonic pricing from environmental economics (Rosen 1974), event studies from finance (Fama et al. 1969), regression discontinuity and modern causal inference from applied microeconomics. The map extends to selected papers in economics, finance, and psychology to ask which method–data–question combinations are productive elsewhere but thinly represented in real estate economics.

In total, 1,499 research ideas are generated under eight conditions and located in that space. The conditions vary in what context the model receives: nothing, the full REE corpus, individual cluster seeds, methods drawn from economics and finance, methods drawn from psychology. Each idea is evaluated on atypicality and a calibrated citation-prediction model, asking what different forms of scaffolding add over a naive prompt.

The results are, on balance, encouraging. The best ideas, particularly those generated through method transfer from economics and finance, score comparably on the validated citation-predictive criterion to the median published paper in the field. Not all AI-generated ideas are good, though. Some land squarely on papers published twenty years ago, having rediscovered questions the field has already answered. But that is also true of human research proposals. The arrival of the AI co-investigator is closer than expected.

2 Related literature and research gap

2.1 AI in economics, finance, and adjacent fields

AI and ML have entered economics and finance as powerful research technologies. Athey and Imbens (2017) argue that ML methods hold promise for improving the credibility of causal identification and policy evaluation; their 2019 survey maps the specific tools — regularisation, random forests, matrix completion, and hybrid ML–econometric estimators — that empirical economists need to adopt. In finance, Gu, Kelly and Xiu (2020) provide a benchmark illustration of what happens when machine learning is brought into an established empirical domain: improved predictive performance, new forms of non-linearity, and a clearer distinction between prediction and economic mechanism. Kelly and Xiu

(2023) survey this financial machine learning literature more broadly, mapping the methods and open questions that remain. Bibliometric mapping of AI's diffusion across economics more broadly — clustering thousands of articles by method and topic — confirms that the wave arrived first in prediction and macroeconomic modelling before spreading to more structural and theoretical applications (Bahoo, Cuñado and Gupta, 2025).

Improved prediction, however, does not distribute its gains evenly. Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022) study the adoption of ML in U.S. mortgage credit screening using roughly 10 million originations. Their key result is that switching from a simple model to a more flexible one is a mean-preserving spread on predicted default probabilities: accuracy rises in aggregate, but there are always borrowers who are reclassified upward and borrowers reclassified downward — winners and losers. Who falls where is not random. Black and Hispanic borrowers are disproportionately less likely to gain from the ML upgrade, partly because the more flexible model better recovers structural relationships that disadvantage them and partly because it more effectively triangulates excluded group characteristics from permissible inputs. The paper establishes that the distributional consequences of deploying AI in high-stakes allocation problems are a first-order empirical concern, not a secondary consideration.

These literatures are relevant here for two reasons. AI typically enters a field first as a measurement or forecasting tool; and they offer mature method–data combinations that can be compared against the real-estate research space for possible import.

2.2 Real estate finance and economics, machine learning, and AI as method

Within real estate finance and economics, the machine-learning literature has grown quickly, particularly around valuation and prediction. Tekouabou et al. (2023) survey AI-based methods for urban real-estate prediction and report systematic predictive gains relative to traditional hedonic approaches. Recent papers in REE and JREFE illustrate a broader shift. Lindenthal and Johnson (2021) use machine learning to classify architectural styles and link them to prices. Lorenz, Kok and co-authors (2023) develop interpretable machine learning for real estate market analysis. Leow and Lindenthal (2025) extend the Gu, Kelly and Xiu (2020) machine-learning asset-pricing programme to REITs: applying ML to traditional factor-return data, they show substantial forecast improvements over OLS — a clean example of AI improving the analysis of conventional data without changing the underlying research question. Calainho, van de Minne and Francke (2024) show the same pattern for commercial real estate price indices: replacing linear hedonic models

with ML improves out-of-sample prediction accuracy on 30,000 New York transactions, while leaving the fundamental index construction problem unchanged.

A separate strand treats AI itself as the object of study. Wan and Lindenthal (2023) develop a framework for testing and validating machine learning systems in real estate. This is meta-research on AI, not research with it. More recently, REE has also published work on combining machine learning and econometrics in commercial real estate and on explainable spatial machine learning for hedonic modelling. Beyond prediction, several papers demonstrate AI unlocking measurements that were previously impossible at scale, enabling traditional research questions to be answered with new data. Naik, Raskar and Hidalgo (2016) use computer vision on street-level imagery to construct city-scale measures of urban appearance quality, applied to a classic question in urban economics about neighbourhood change. Shen and Ross (2021) apply unsupervised NLP to MLS listing descriptions, extracting a description-uniqueness measure that captures soft information about property quality, and find that a one-standard-deviation increase raises sale prices by around 15 percent. Bartik, Gupta and Milo (2025) extend this logic to regulatory text: their generative regulatory measurement approach reads thousands of municipal zoning codes and extracts structured indicators at near-human accuracy, enabling a national analysis of housing regulation that would otherwise be infeasible.

These papers establish that AI adds value in real estate research. The gains are concentrated in measurement, prediction, feature extraction, and model support: RA tasks. Whether AI can do more is the question this paper pursues.

2.3 Bibliometrics and the science of science

Donthu et al. (2021) survey bibliometric methods: citation analysis, co-authorship networks, and co-word mapping. Fortunato et al. (2018) situate such work within the science-of-science programme, which studies the structure and dynamics of knowledge production itself.

The most directly relevant work concerns the relationship between novelty, combination, and impact. Uzzi, Mukherjee, Stringer and Jones (2013) show that high-impact papers tend to combine a mostly conventional base of prior work with a small number of atypical combinations — pairings of knowledge that are unusual but not unprecedented. This empirical finding provides a bibliometric analogue to the incremental/recombinant/frontier distinction: high-impact work is not purely incremental (which lacks novelty) nor purely frontier (which lacks legibility), but sits at the boundary between the familiar and the new. Foster, Rzhetsky and Evans (2015) extend this line of inquiry to research strategies,

documenting how scientists trade off the higher expected rewards of exploratory research against the lower risk of incremental work, and showing that conservative strategies dominate even when novel research yields larger returns when it succeeds.

A second relevant strand uses semantic embeddings to characterise research spaces rather than citation networks. Tshitoyan et al. (2019) demonstrate that word embeddings trained on materials science abstracts encode implicit knowledge about element combinations and can predict future discoveries — papers not yet written — by identifying concepts that are close in embedding space but not yet combined in the literature. This is a direct precursor to the approach taken here: the embedding space is not merely a visualisation of what has been done, but a structure that encodes what is possible at the frontier.

This paper differs from these studies in two ways. First, the focus is on the research content of papers rather than metadata or citation proxies. Second, the objective is not only to describe the research space but to use it as an empirical benchmark against which new proposals can be located — asking where a given idea falls relative to the existing frontier, not just whether it is novel in a statistical sense.

2.4 LLM use in academic writing, scientific discovery, and idea generation

A parallel literature documents how researchers themselves use AI tools. Kobak et al. (2025) introduce an excess-word method for detecting LLM involvement in scientific abstracts, tracking words that language models overuse relative to human writers. Applied to 14 million biomedical abstracts, they estimate that at least 13.5 per cent of 2024 papers were processed with an LLM. Their method provides a lower bound — authors who edit LLM drafts carefully leave no trace — and a reproducible empirical benchmark. Section 5.1 replicates the analysis for the real-estate literature.

A growing literature asks whether LLMs can contribute to scientific discovery beyond summarisation and coding. Si, Hashimoto and Yang (2025a) provide the most relevant benchmark: a large-scale human evaluation of AI-generated research ideas in NLP, finding higher judged novelty but lower feasibility, alongside evidence that generation systems may still lack diversity. A follow-up study (Si, Hashimoto and Yang, 2025b) examined what happens when the ideas are actually implemented: NLP researchers spent up to three months building the AI- and human-suggested projects and then had independent assessors rate the outcomes. Ideas that had seemed more novel and exciting before execution were rated worse after it, while human-generated ideas improved. The "ideation-execution gap"

— AI-generated ideas look better at the proposal stage than at the results stage — is a direct challenge to any evaluation method that stops at ideation. In the natural sciences, the execution step can at least be partially validated: a drug candidate either works in the animal model or it does not; a predicted material property either matches the synthesis or it does not. In social science and economics, no equivalent bench test exists. The field cannot wait years for a paper to be executed, published, and cited before judging whether a proposal was worth pursuing. This asymmetry motivates the approach taken here: using the structure of the field's existing literature — specifically atypicality, which retrodictively predicts citations — as a measurable ex-ante proxy for scientific value.

More recent agentic systems push further towards autonomous idea generation, refinement, and evaluation: Su et al. (2025) propose a multi-agent system (VirSci) in which specialised LLM agents collaborate to generate and critique ideas, while Baek et al. (2025) develop ResearchAgent, which iteratively refines proposals by querying the scientific literature. Gottweis et al. (2025) describe Google's AI co-scientist, a multi-agent system built on Gemini 2.0 that generates, debates, and evolves hypotheses using a tournament process; early results in biomedicine include experimentally validated drug-repurposing candidates and an independently reproduced discovery in bacterial gene transfer. OpenAI's GPT-Rosalind (OpenAI 2026), released in April 2026, is a frontier reasoning model purpose-built for biology, drug discovery, and translational medicine, designed to support "evidence synthesis, hypothesis generation, experimental planning, and other multi-step research tasks" across the early stages of scientific discovery. Lu et al. (2026) close the loop entirely: their AI Scientist generates ideas, writes code, runs experiments, produces manuscripts, and performs automated peer review — a paper it generated passed the first round of review at a top machine learning workshop. They identify evaluation as the binding constraint: "A central challenge in developing such a system is automatically evaluating the quality of its scientific output at scale." That is the same problem this paper addresses, in a field where automated execution does not exist. These papers matter here not because their settings transfer mechanically to real estate economics, but because they clarify the question this paper tests: whether AI can move from assistance to ideation in a disciplined, measurable way — and what the limits of that move are when no wet-lab equivalent exists to adjudicate the outcome.

2.5 Novelty, research strategy, and the frontier

Uzzi et al. (2013) establish that the most-cited papers tend to combine conventional and atypical knowledge in specific proportions: genuine advances are rarely purely incremental or purely sui generis. Foster et al. (2015) show that scientists systematically underinvest

in novel research relative to the social optimum, preferring incremental strategies that reduce the risk of failure. Wang, Veugelers and Stephan (2017) document that truly novel papers are initially under-cited and face systematic resistance from peer review, suggesting that the research space is not a level playing field: ideas that are genuinely distant from the current frontier face higher barriers to publication and recognition.

These findings shape how to interpret the proposal-location results. A proposal in a sparse or frontier region is not guaranteed to be valuable — it may simply be unfamiliar. But the science-of-science evidence suggests the expected payoff, conditional on acceptance, is higher than for incremental work. Embedding distance is therefore informative about risk and return, not just novelty.

Tshitoyan et al. (2019) add a further dimension: the embedding space trained on a corpus encodes not only what has been done but what the literature is ready to support. Concepts that co-occur frequently but have not yet been combined in a single paper represent low-hanging combinations — the background knowledge exists, only synthesis is needed. The most productive frontier proposals are therefore not those at maximum distance from existing work but those bridging dense clusters: recombinant rather than purely novel.

2.6 Methods imports and cross-field transfer

Real estate finance and economics has repeatedly advanced by importing methods from neighbouring literatures. Hedonic pricing comes from environmental and labour economics; event studies from finance; regression discontinuity and modern causal inference from applied microeconomics. This history motivates the cross-field transfer exercise. Rather than asking only what is missing within real estate, the exercise asks which combinations of methods, data, and questions are already productive in other fields — finance, economics, psychology, and allied data-rich social sciences — but have not yet been systematically brought into real estate economics.

At the same time, not every transfer is valuable. Prediction methods transplanted into causal settings without regard to identification, institutional context, or the data-generating process do not necessarily advance economic knowledge. The purpose of the cross-field comparison is therefore not simply to import novelty, but to identify imports that satisfy the inferential standards of the receiving field.

2.7 Research questions

Two linked questions organise the paper.

RQ1. How deeply has AI penetrated real estate research — from writing assistance to methodological enablement — and in what roles? AI enters research at multiple levels, not all of them visible. At the surface, excess-word analysis reveals LLM involvement in drafting abstracts. Below that, keyword detection identifies papers where AI performs tasks that would otherwise be infeasible — classifying images, parsing regulatory text, fitting high-dimensional models. Between these two sits a grey zone of AI use for data collection, literature screening, and other conventional support work that leaves no identifiable trace. RQ1 documents the pattern across levels: how much AI has entered the field, at what stage of the research process, and in what role.

RQ2. Can AI move from research assistant to principal investigator? Can AI do more than executing well-defined tasks within research designs conceived by human researchers? Can it formulate questions worth asking, identify combinations of methods and data that the field has not yet tried, and generate proposals that land in genuinely new territory rather than replicating what already exists.

3 Data

3.1 Text as data

Gentzkow, Kelly and Taddy (2019) provide the canonical overview of text as data in economics. The use of large language models here extends this tradition: structured summaries are extracted from each paper's full text, covering research question, method, data source, sector, and main finding. This identifies substantive similarities between papers that use different surface vocabularies, and permits direct comparison along dimensions that matter for economic research — not only what topic is studied, but how, with what data, and to what inferential end.

The empirics in this paper rely on the universe of papers published in core real estate journals:

Real Estate Economics (REE) — formerly the *Journal of the American Real Estate and Urban Economics Association* — is the field's oldest specialist journal, published continuously since 1973. Metadata for all 1,676 articles (1973–2026) were collected via the CrossRef API. Full-text PDFs were obtained for 909 articles (1997–2026, approximately 75% of the post-1996 archive) via the Wiley Text and Data Mining API under institutional

licence; pre-1997 papers are not available through this channel.

Journal of Real Estate Finance and Economics (JREFE) — published by Springer since 1988, covering real estate finance, investment, and urban economics. Metadata collected for 1,702 articles via OpenAlex. Full-text access via Springer TDM API.[1]

Journal of Urban Economics (JUE) — published by Elsevier since 1974. Broader in scope than the specialist journals but a major outlet for housing supply, land use, and spatial economics. Of the full archive, only the subset whose OpenAlex topics include housing markets, land use, urban development, valuation, or closely related areas is retained.

3.2 Supplementary journal coverage

A targeted subset of real estate, housing, mortgage, and urban papers is drawn from leading general-interest and neighbouring journals. The economics and finance set includes the *Journal of Finance*, *Review of Financial Studies*, *Journal of Financial Economics*, *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*. The supplementary sample is not intended to provide exhaustive coverage of those fields. Its purpose is to support cross-field transfer analysis by identifying method–data–question combinations that are productive elsewhere but thinly represented in real estate finance and economics.

For the psychology component, two journals are drawn on. *Psychological Methods* is where methodological innovation in psychology is published and debated: structural equation models, causal inference designs, measurement invariance, and computational approaches that typically reach adjacent social sciences a few years later. *Psychological Science* is the empirical flagship of the Association for Psychological Science, setting the discipline's standards for replication, pre-registration, and open science. The two journals cover complementary dimensions of cross-field transfer: new tools from *Psychological Methods* and new questions about behaviour, judgment, and decision-making under uncertainty from *Psychological Science*, the latter directly applicable to housing choice, mortgage decisions, and investor behaviour.

[1] https://dev.springernature.com/

3.3 Citation and metadata enrichment

Citation counts, reference lists, concept tags, and journal metadata are obtained via OpenAlex.[2] Where full text is unavailable, title, abstract, and reference information are retained for cross-field comparisons.

3.4 Corpus roles

In this paper, REE is used as the primary field-defining corpus: it sets the research space that all other journals are positioned relative to and provides the base case for the idea-generation experiment. JREFE and JUE extend coverage to the specialist finance and urban economics literatures that share substantial overlap with REE but have distinct methodological emphases. The economics and finance journals (AER, JF, RFS) supply the cross-field transfer pool: papers selected for methodological distinctiveness from the real estate literature, providing the source material for conditions D and F in the generation experiment. The psychology journals serve the same transfer role for behavioural methods and questions. REE is the natural base case for this analysis given its scope and continuity, but the design is not specific to it: any sufficiently large specialist journal corpus could occupy the same role, and a replication using JREFE or JUE as the primary field would test whether the patterns identified here are features of the discipline or of the particular outlet.

4 Method

4.1 Structured extraction via LLM

For each paper with a full-text PDF, the text is extracted and an LLM queried[3] with a structured prompt to recover the following fields:

- `research_question`: Main research question or objective,
- `method`: Primary empirical or theoretical method,
- `data`: Main data source, geography, and period,

[2] OpenAlex is a free, fully open bibliometric database maintained by OurResearch (https://openalex.org), covering over 250 million scholarly works with metadata, abstracts, citation networks, and concept tags.

[3] All extractions use Claude Sonnet (Anthropic). Spot-checking a 50-paper sample against GPT-4o produces near-identical structured outputs; the task is constrained enough that the choice of frontier model does not materially affect results. This may need revisiting if model capabilities or instruction-following behaviour diverge substantially in future generations.

- `sector`: Sector(s) studied,
- `ai_ml_role`: Role of AI or machine learning, if any, and
- `key_finding`: Main contribution or result.

The extraction pipeline is cached incrementally so that it is fully resumable.[4]

4.2 Validation of extraction

The whole framework rests on extraction quality, so validation is not a footnote. A small sample across journals, periods, and topic areas is hand-coded, with manual coding compared against LLM outputs. Agreement is assessed separately for method, data, sector, and AI-role classification. Common failure modes are also investigated — notably confusion between prediction and causal analysis, and between core and auxiliary datasets — to establish that the research-space representation captures the underlying economics of papers rather than their surface language. As many students will happily claim, LLMs can indeed summarise the vast majority of complex academic papers accurately.

4.3 Embedding, clustering, and interpreting the space

The substantive extraction fields are concatenated into a single text representation per paper and encoded using a sentence-transformer model, generating one embedding per paper. All journals are embedded in a shared space so that distances are directly comparable across outlets and fields.

Dimensionality reduction is used for visualisation and clustering, with clustering performed in a higher-dimensional reduced space rather than in two-dimensional plots. Clustering uses HDBSCAN, a density-based algorithm that identifies clusters as regions of high local density and leaves low-density points unassigned rather than forcing every paper into the nearest cluster. This is a deliberate choice: papers that sit between research communities, combine unusual method–topic pairs, or simply resist clean categorisation are better left unassigned than absorbed into a cluster they do not belong to. The 870 unassigned papers (19.5% of the corpus) are not a failure of the method but a feature of

[4]The system prompt used for all extractions reads in full: *"You are a research assistant analysing academic real estate economics papers. Given information about a paper, extract exactly five fields as JSON:* `research_question` *— one sentence: the main research question or objective;* `method` *— one sentence: the primary empirical or theoretical method;* `data` *— one sentence: the main data source(s) and coverage;* `key_finding` *— one sentence: the main result or contribution;* `ai_ml_role` *— one sentence: the role of AI/ML in this paper, or 'None' if absent. Be concise. Use plain English. Do not copy abstract text verbatim."* The user message appends the paper title, abstract, and a full-text excerpt (up to approximately 8,000 tokens) where the PDF is available, or title and abstract only where it is not.

the research space — they mark territory that is genuinely diffuse rather than clustered. Cluster labels are then generated from representative papers and refined by human inspection. The resulting clusters define the major research strands in the literature and yield an empirical typology of what the field has actually done. Dense established clusters mark ground where new work is likely to be incremental; sparse bridging regions hold recombinant combinations; frontier regions contain topic–method–data configurations that appear largely unrepresented. This makes the research space usable as an evaluation device for new ideas.

4.4 Locating proposals in the research space

Research proposals are generated with LLMs in multiple ways — with different contexts, prompts and cross-field transfers from adjacent disciplines — and embedded into the shared space. Each proposal is expressed in a structured form: research question, empirical design, likely data source, expected contribution, and relevance to real estate economics. Incremental proposals fall inside existing dense clusters; recombinant proposals lie in sparse space between clusters; frontier proposals lie at or beyond the observed support of the literature.

This is a measure of **distance from existing research configurations**, not a complete measure of innovation. Distance does not imply importance, correctness, or feasibility — but it provides a disciplined, replicable way to compare whether ideas are more of the same or genuinely different. The question is not which source produces better ideas — that is a matter of judgement the coordinate system cannot resolve. The question is descriptive: where do proposals from different sources tend to land, and does the form of scaffolding change that distribution?

4.5 Cross-field comparisons

Papers from the *Journal of Finance*, *Review of Financial Studies*, *Journal of Financial Economics*, *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* are embedded alongside the specialist-journal corpus and their cluster positions compared. The comparison is descriptive. Publication venue is not a clean quality signal: papers that are innovative in ways specific to real estate, or that rely on sector-specific data and institutional context, may never enter the top-journal selection process at all. Real estate papers in those journals are drawn from the intersection of high quality and legibility to a general-audience editor, not from the full frontier of the field. Any observed differences in cluster position have multiple plausible explanations —

method origin, research fashion, genuine novelty, or simply high execution of approaches already common in specialist journals — and journal type is treated as one covariate among several alongside citation counts, year, and topic.

For cross-field transfer, Comparable embeddings are constructed for selected papers in finance, economics, psychology, and related disciplines, and clusters or local configurations that are productive there but underrepresented in real estate economics are identified. The key safeguard is economic validity: imported methods are only counted as promising when the underlying inferential logic matches the target real-estate question. A purely predictive architecture imported into a causal policy setting does not qualify.

5 Results

5.1 LLM use in real estate research writing

Kobak et al. (2025) identified a set of words that large language models overuse relative to human writers — "delve," "pivotal," "meticulous," "nuanced," "comprehensive," "crucial" — and showed that their frequency in biomedical abstracts jumped sharply after ChatGPT's November 2022 launch. The same excess-word method is applied to two real-estate corpora. The first is the full *Real Estate Economics* abstract archive (up to 66 abstracts per year from 2010 to 2026); the second is approximately 101,000 papers indexed by OpenAlex under real-estate, housing, urban economics, and REIT concepts from 2010 to 2025.

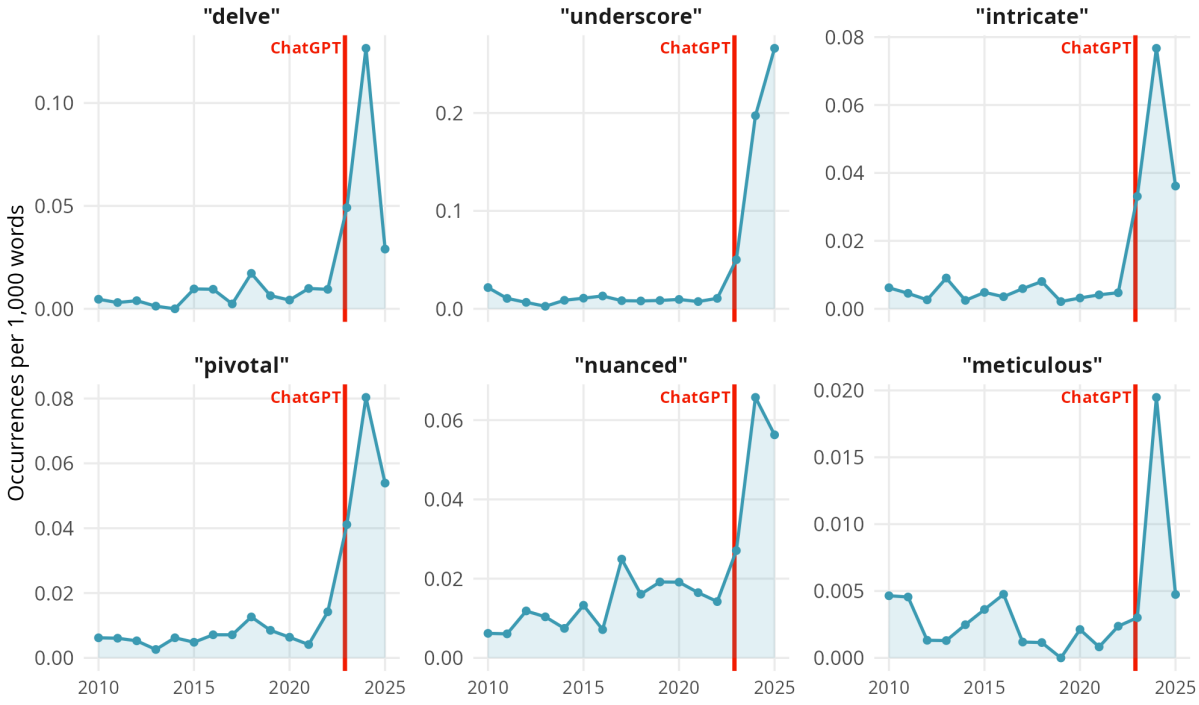
The REE corpus is too small for year-level inference — the word counts are sparse even for common words — so it serves only to confirm the direction; the OpenAlex sample provides the quantification. In that broader sample, aggregate LLM-indicator word rates were stable from 2010 to 2022, averaging roughly 0.85 occurrences per 1,000 abstract words (Figure 2). In 2023 the rate rose to 1.31 per 1,000, and by 2024 it reached 2.45 per 1,000 — a three-fold increase in two years. Individual words are sharper. "Delve" was essentially absent before 2023 (0.006/1k averaged across 2010–2022), appeared at 0.049 in 2023, and reached 0.127 in 2024, a twenty-fold increase. "Pivotal" rose eleven-fold over the same horizon; "crucial" four-fold; "comprehensive" three-fold. These patterns replicate Kobak et al. (2025) in a field with no prior documentation of LLM adoption in writing.

Two interpretive cautions. First, excess-word rates provide a lower bound on LLM involvement: authors who use AI for drafts and then edit carefully will not leave a detectable signature. The true adoption rate is higher than the word-frequency evidence suggests. Second, some flagged words — "robust," "leverage" — have legitimate technical uses in economics and finance that are unrelated to LLM-assisted writing; they are

excluded from the trend analysis. The words driving the 2023–2024 jump — "delve," "pivotal," "nuanced" — are the clean diagnostic indicators that Kobak et al. (2025) identify as having essentially no pre-ChatGPT baseline in scientific writing.

The vocabulary is now leaking into human spoken communication. Yakura et al. (2025) apply causal inference techniques to 740,000 hours of transcribed speech — 360,000 YouTube academic talks and 770,000 podcast episodes — and detect a sharp, abrupt increase in LLM-preferred words in human oral language after ChatGPT's release. Machines trained on human writing are now reshaping what humans say aloud: a closed feedback loop with uncertain implications for where the linguistic signal goes next.

Figure 1: LLM-Indicator Word Rates in Real Estate and Housing Research, 2010–2025



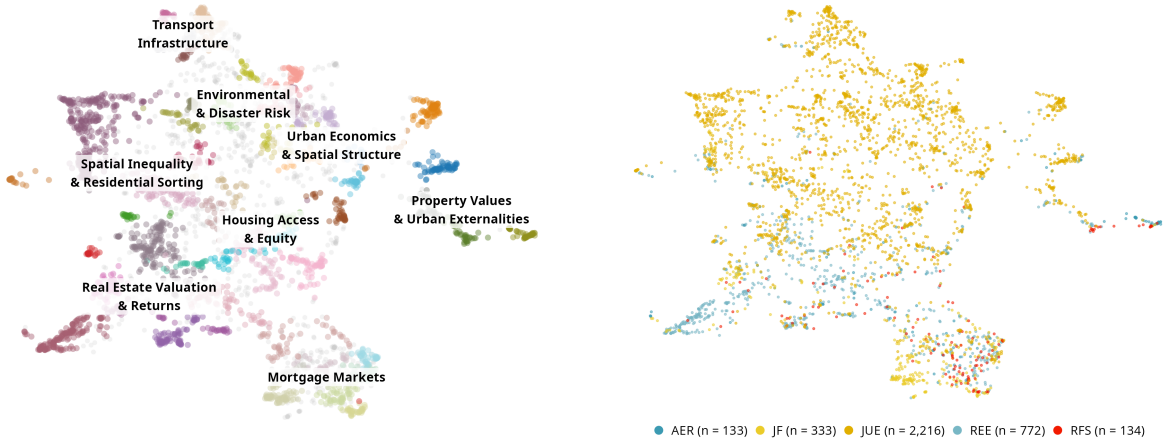
Notes: Annual rate (occurrences per 1,000 abstract words) of fifteen LLM-associated words identified by Kobak et al. (2025) in approximately 101,000 papers indexed by OpenAlex under real-estate, housing, urban economics, and REIT concepts. Rates are aggregated across all indicator words. The dashed vertical line marks November 2022 (ChatGPT launch).

5.2 The research space of real estate economics

The combined embedding space (3,591 papers after removing administrative content) can be sorted into 51 clusters (Full table in the appendix, sorted by cluster size). Each cluster is labelled by a short topic description generated by an LLM from a sample of paper titles

and research questions. The UMAP visualisations below show the research space coloured by cluster and by journal (Figure 1).

Figure 2: Research Space of Real Estate Finance and Economics



(a) Research clusters

(b) Journal of publication

Notes: All 3,591 papers embedded using the same sentence transformer. (a) Coloured by HDBSCAN cluster; meta-cluster labels positioned at cluster centroids. (b) Coloured by journal of publication. Each point is one paper; position reflects semantic similarity of research question, method, and context.

5.3 The emergence of AI

AI and machine learning papers within the REE corpus are identified by searching the LLM-extracted method and finding fields for a set of method-specific terms — machine learning, neural network, random forest, gradient boosting, natural language processing, computer vision, word embeddings, textual analysis, and related — and manually verifying each match. This yields 12 confirmed AI/ML papers among 772 REE articles with full-text extraction (roughly 1.6 per cent of the full-text sample, but 4–7 per cent of annual output in recent years). The count is a lower bound: it captures only papers where AI performs a task that was otherwise infeasible and where the method is explicit enough to appear in the extraction. Papers that used AI for data collection, source screening, or other support tasks that remain invisible in the methods section are not counted. The Kobak analysis in Section 5.1 captures part of this grey zone at the writing stage; the empirical middle — AI used to do traditional research tasks without either being identified as ML or leaving a linguistic trace in the abstract — is not directly observable.

There are no AI/ML papers in REE before 2020; isolated examples appear in 2020 and 2022. The first concentration emerges in 2023, when four ML papers appeared in a single year — 7 per cent of that year's output. The share runs at 4–6 per cent in

2024–2026. This is a post-pandemic phenomenon: whatever adoption was happening in adjacent fields (Gu, Kelly and Xiu published in RFS in 2020; Bartik, Gupta and Milo's regulatory measurement work circulated from 2022), REE did not publish AI-methods papers at a detectable rate until 2023.

The adoption is heavily concentrated. Of the 12 papers, 5 (42 per cent) belong to cluster 21 — hedonic pricing and residential location choice. This is the natural entry point: ML improves the nonparametric estimation of the price function without requiring model specification, and data are abundant and structured. The remaining 7 papers span 6 named clusters — transit capitalisation, smoking and rental markets, housing dynamics, asset valuation, human mobility, and REIT forecasting — with one paper in the unclustered noise class. AI papers do not form a methodological island but appear embedded within the existing topic clusters where the methods are being applied (Figure 3).

Most papers that employ AI use it as a *measurement or prediction tool*. Three examples illustrate the range.

Lorenz, Willwersch, Cajias and Fuerst (2023) apply interpretable machine learning to the hedonic estimation of residential rents, comparing XGBoost with model-agnostic interpretation methods against a conventional hedonic specification. Size and age emerge as the dominant rent drivers, while interactions — large apartments in historic buildings with balconies in affluent neighbourhoods — attract disproportionate premia that a linear model would miss. The paper uses ML to improve measurement precision while retaining economic interpretability, a design that has become the template for hedonic ML applications.

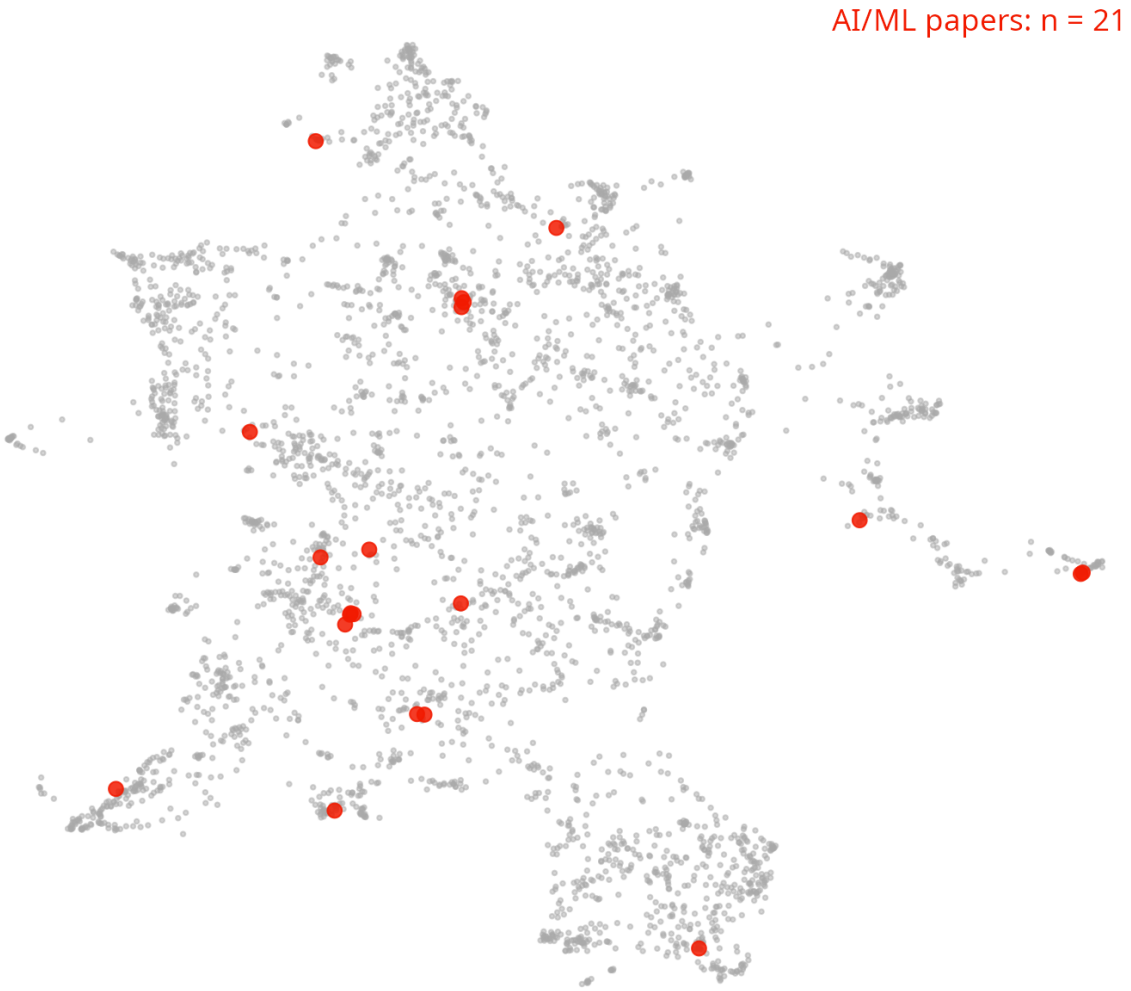
Francke and van de Minne (2024) address the analogous problem in commercial real estate, where small samples and high property heterogeneity make conventional hedonic models unstable. They propose an iterative hybrid: a random effects model captures common trends, property-type trends, and location values, while a separate ML algorithm fits the nonlinear relationship between property characteristics and price. Applied to 2,652 commercial transactions in Phoenix between 2001 and 2021, the hybrid achieves an out-of-sample mean absolute percentage error below 11 per cent, lower than either component alone.

Lopez, McCoy and Sah (2022) use ML differently: to extract structured information from property listing data and identify anticompetitive behaviour. Agents representing foreclosure sellers frequently require buyers to obtain pre-qualification letters from seller-preferred lenders even when buyers are already pre-qualified with independent lenders; more than half of the steering activity originates from foreclosure property sellers. ML here

is a text-mining device that makes a novel institutional fact observable, not a forecasting tool or a hedonic estimator.

None of the identified 12 papers uses AI for causal inference support or studies AI as an economic phenomenon — the two roles most prevalent in the neighbouring economics and finance journals. REE's AI/ML papers are uniformly at the measurement and prediction end of the spectrum: AI as a tool for doing something the researcher has already specified, not as a co-designer of the research itself.

Figure 3: AI/ML Papers in the Research Space



Notes: Grey dots = all corpus papers (n = 3,593, umap_x < 15). Red dots = AI/ML-classified papers (n = 21, 12 within REE). Each point is one paper; position reflects semantic similarity in the shared embedding space.

5.4 Real estate papers in economics and finance journals

The combined embedding space includes 2,898 papers from REE/JREFE/JUE and 572 papers drawn from the real-estate-relevant subsets of AER, JF, and RFS. Because all papers are embedded using the same LLM-extracted structured representations and the same sentence transformer, the resulting clusters reflect content similarity rather than source. A paper that asks the same question using the same method as a REE article will land in the same neighbourhood regardless of where it was published.

Shared versus exclusive territory. Of the 51 clusters, 37 contain papers from both source types; 14 contain only top-journal papers; none contains only REE/JREFE papers. That is: the research space of REE and its sister journals is a strict subset of the space occupied by the broader real-estate literature. Every region where REE publishes is a region where top journals have also published; there are regions where top journals publish but REE does not.

The 14 exclusively top-journal clusters are dominated by JUE papers (78 per cent of the 572 papers in those clusters), covering: congestion pricing and transport capacity optimisation, firm location decisions and agglomeration economies, spatial labour markets and worker mobility, climate risk and financial markets, urban structure and spatial growth patterns, city size distribution, residential mobility and migration, urban density modelling, spatial mismatch and racial employment inequality, gentrification and neighbourhood change, and urban/spatial equilibrium theory. These are active areas of urban economics and finance that REE does not engage at all — not because the topics are outside the remit of real estate economics, but because they are developed in the broader economics and finance literature without finding their way back into REE. Climate risk and financial markets (cluster 7, n = 44, median 138 citations) is the most striking example: a fast-growing area with large citation counts that has a substantial presence in JF and AER but no equivalent cluster in REE.

REE arrives later in shared territory. Among the 37 mixed clusters, papers from AER, JF, and RFS have a mean publication year of 2000.3, compared with 2015.0 for REE papers in the same clusters. The 15-year gap reflects two distinct patterns. First, some topics were established first in the general economics literature (land use regulation, local tax policy, racial segregation in housing) and subsequently picked up by REE. Second, some recent REE papers enter clusters that top-journal papers anchored two to three decades earlier, where the question has arguably been settled fully but REE is still active.

Citation gradient within REE. Within the REE corpus, there is a systematic relationship between cluster composition and citations. REE papers in REE-dominated

clusters (fewer than 25 per cent top-journal peers) have a median citation count of 28; those in top-journal-dominated clusters (more than 75 per cent top-journal peers) have a median of only 9. The mean publication year for REE papers in top-dominated clusters is 2018.6, compared with 2013.3 in REE-dominated clusters, so part of the gap reflects youth. However, the age-adjusted citation residual remains significantly lower for REE papers in mixed clusters relative to top-journal papers in the same clusters (mean residual -0.40 vs. $+0.17$, $p < 0.0001$). This pattern is consistent with the well-documented advantage of publishing in higher-ranked outlets — top-journal papers attract more citations even on the same topic — but it may also reflect self selection of authors at submission, reflecting paper quality and scholarly incentive structures.

The research space boundary. The absence of REE-only clusters has a simple interpretation: REE does not occupy any research territory that is completely absent from the broader economics and finance literature. This is simultaneously reassuring (the field is intellectually coherent with its neighbours) and limiting (it may suggest that REE follows methodological and topical trends set elsewhere rather than originating them). The cross-field transfer analysis in Section 5.5 takes this boundary as a starting point: if the research space of real estate economics is a subset of the broader space, the question is which regions of the broader space are densely populated elsewhere but thinly represented in REE — and whether AI-assisted method transfer can help close the gap.

5.5 AI as PI: generating and evaluating research proposals

The preceding sections document AI extracting information, classifying documents, and measuring the structure of the literature. This section asks whether AI can act as a principal investigator: generating proposals that are not merely plausible-sounding but meaningfully novel.

Ideas are generated under eight conditions, evaluated using atypicality as the primary criterion, comparing what different forms of embedding-space scaffold add over a naive prompt. Each condition is run with and without a hard data-existence constraint, yielding sixteen generation runs in total.

5.5.1 Experimental design

Ideas are generated under eight conditions using the same large language model. Each condition is run twice: once unconstrained, and once with a hard data-existence requirement embedded in the prompt (*"only propose designs where you can confirm the required data exists and is accessible to an academic researcher"*). This yields fourteen generation

runs in total and allows me to test whether the constraint changes what gets proposed or merely strips out ideas that a post-hoc feasibility assessment would have flagged anyway.

A — Naive. No corpus context. The AI operates on training priors alone, generating 100 ideas in four batches of 25.

B — Full REE corpus. All 786 extracted REE paper summaries are passed as context. To guard against position effects in long contexts — where the model may attend disproportionately to papers near the start or end of the input — the order of summaries is randomised before each batch of 25 ideas, so the context is reshuffled four times across the 100 ideas. Anthropic's prompt caching avoids reprocessing the 110k-token context on every call: the shuffled prefix is cached at the start of each batch, reducing the per-call input cost for subsequent calls within that batch to roughly $0.03. The cache is invalidated and rebuilt on each reshuffle.

C — Cluster-seeded. One prompt per cluster, supplying the cluster label and two to three representative paper summaries from that cluster. One idea per cluster $\times$ 51 clusters = 51 ideas.

D_econ — Econ/finance method transfer. Both the REE corpus (cached, reshuffled as in B) and approximately 150 papers from economics and finance are passed as context. The source-field papers are selected not at random but for maximum methodological distinctiveness: specifically, papers falling in clusters where REE representation is below 10 per cent. The prompt instructs the model to propose real estate research designs that apply methods from the second set to questions in the first, with the explicit requirement that the inferential logic of the method must transfer cleanly to the real estate setting.

D_psych — Psychology method transfer. Identical structure to D_econ, with the source-field context replaced by ~150 methodologically distinctive psychology papers (from *Psychological Methods* and *Psychological Science*, selected on the same low-REE-overlap criterion).

E_econ — Econ/finance question and method transfer. As D_econ, but the prompt asks the model to transplant both the research question and the empirical design from the source-field papers, not only the method. This tests whether importing a complete study logic — question, design, and expected mechanism — produces ideas of different quality than importing method alone.

E_psych — Psychology question and method transfer. As D_psych, with full study transplant rather than method import only.

All ideas share a common output structure: research question, empirical method,

required data, expected contribution, a self-rated feasibility score (0–10), and a novelty rationale in which the model explains why it considers the proposal distinct from existing work. The novelty rationale is used to assess claimed versus actual innovation: A structured novelty score (1–4 scale: 1 = replication; 2 = extension; 3 = new method or dataset; 4 = new question) is extracted from the rationale text in a separate call and compared against the idea's geometric position in the embedding space. This tests whether the model is a reliable reporter of its own novelty — and, critically, whether providing more context (conditions B, D, E) reduces systematic overclaiming relative to the context-free baseline (A).

F — Citation-signal corpus. As condition B, but each REE paper summary is augmented with its age-adjusted citation residual — the OLS residual from regressing log(1 + citations) on publication year, which strips out the mechanical age effect and leaves a clean within-vintage impact signal. The prompt instructs the model to use these signals when forming its proposals: *"Papers with higher citation scores were judged more impactful by the field. Generate ideas you expect to receive high future citations."* This tests whether explicit impact feedback changes what the model proposes, and whether directing it toward citation impact produces ideas that score higher on atypicality, feasibility, or both — or whether the model conflates citation signal with methodological conservatism and regresses toward well-cited but incremental territory. The same caching and reshuffling protocol as condition B applies.

The scaffold is the treatment; the embedding pipeline, atypicality measure, and citation-prediction regression are held constant across all conditions.

Context and caching protocol. Conditions B, D, E, and F all pass the full REE corpus (786 summaries, ~110k tokens) as the prompt prefix. This prefix is cached using Anthropic's prompt caching to avoid reprocessing 110k tokens on every call; the per-call cost for a cached prefix drops from ~$0.33 to ~$0.03. To guard against position effects — where the model attends disproportionately to papers near the start or end of a long context — the order of summaries is randomised before each batch of 25 ideas and the cache is rebuilt on each reshuffle, yielding four distinct orderings per condition. For conditions D and E, the ~150 source-field papers are appended after the cached REE prefix; only those additional ~25k tokens are billed at the full input rate per call. Total estimated API cost across all sixteen runs (eight conditions × two constraint variants) is approximately $45–50 on Claude Sonnet.

Data-existence constraint. Every condition is run twice. In the unconstrained variant the model generates freely. In the constrained variant the prompt adds: *"Only propose designs where you can confirm the required data exists and is accessible to an*

academic researcher." Comparing the two variants within each condition tests whether the constraint changes what gets proposed or merely filters ideas that a post-hoc feasibility assessment would have removed anyway.

Claimed versus actual innovation. Each idea includes a free-text novelty rationale in which the model explains why it considers the proposal distinct from existing work. A structured novelty score (1–4, matching the same scale applied to published REE papers) is extracted from this rationale in a separate call and compared against the idea's geometric position in the embedding space. The hypothesis is that models with more context (conditions B, D, E, F) overclaim less — i.e. their self-assessed novelty correlates more closely with actual embedding-space distance from the frontier.

5.5.2 Evaluation and results

Each run produces 100 ideas (51 for cluster-seeded, which produces one idea per cluster), across sixteen generation runs — eight conditions × two constraint variants. Each idea is scored on atypicality and feasibility and assigned to the nearest HDBSCAN cluster in the embedding space.

Generated ideas are evaluated on atypicality — Shannon entropy of the cluster distribution among the idea's 20 nearest neighbours in the embedding space, normalised to [0, 1]. An idea whose neighbours all belong to the same cluster scores near zero; an idea whose neighbours are spread across many clusters scores near one. This operationalises the Uzzi et al. (2013) notion of atypical combination: the idea bridges existing research communities rather than extending a single one.

Atypicality is distinct from novelty (nn_dist — cosine distance to the nearest existing paper), which measures local density: low nn_dist means the idea lands in a densely populated region of the embedding space, close to many existing papers; high nn_dist means it sits in sparse, thinly covered territory. An idea can have low nn_dist and high atypicality — sitting at the busy intersection of several research strands, legible but combinatorially unusual — or high nn_dist and low atypicality — isolated in sparse space but extending a single strand rather than bridging several. The former is the Uzzi et al. sweet spot; the latter is what the citation evidence penalises. Atypicality is also distinct from within-cluster frontier distance (centroid distance), which carries no citation signal once paper age is controlled: being peripheral within a research cluster does not make a paper more impactful, it just makes it newer. Atypicality asks a different question — not whether a paper departs from its own cluster's centre, but whether its neighbourhood spans multiple communities. The choice of atypicality as the primary evaluation criterion

follows the Uzzi et al. (2013) argument: high-impact science combines mostly conventional foundations with a small number of atypical elements, and the citation evidence in the prediction model is consistent with that — nn_dist is a significant *negative* predictor of citations ($\beta = -2.43$, $p < 0.001$), meaning papers that sit far from all existing work attract fewer citations, not more. What the citation model rewards, conditional on proximity, is location in a well-cited part of the research space; atypicality is the ex-ante signal for whether a proposal bridges communities in the way Uzzi et al. identify as productive, computed before cluster placement is known.

Feasibility is handled at the generation stage rather than through a scoring composite. Every condition is run twice: once unconstrained, and once with a hard data-existence requirement embedded in the prompt. Section 5.5.3 reports a structured post-hoc feasibility assessment that verifies whether the constraint does its job. Published papers in the validation sample are all feasible by definition, which means feasibility cannot be estimated as a regression coefficient from the training data; it is a precondition, not a predictor.

Table 2 summarises means by run; Figure 4 shows where ideas land in the research space relative to the existing corpus.

Figure 4 maps where each run's ideas land in the embedding space. The visual pattern is consistent across most conditions: ideas spread across the main body of the research space, with mild concentrations in the largest clusters. The Herfindahl–Hirschman index (HHI, Table 2) makes the concentration differences precise. The REE corpus itself has an HHI of 0.091, reflecting the uneven size distribution across its 52 categories (51 named clusters plus the noise category). Cluster-seeded generation (C) is the most dispersed — HHI ≈ 0.044 — by construction: one idea per cluster forces even coverage. Naïve (A) and citation-guided (H) are the most dispersed among the unconstrained runs: without any cross-field signal, the model spreads proposals broadly and, in H's case, slightly favours the higher-citation regions of the corpus without collapsing onto them.

The psych-sourced conditions concentrate ideas sharply. Psychology paradigm transfer (G, unconstrained) reaches HHI = 0.161 — nearly twice the corpus baseline — because the model gravitates to a narrow behavioural real estate territory where transplanted psych phenomena already have natural homes: loss aversion in selling decisions, anchoring in appraisals, attention effects in search. The data constraint pulls G back to 0.100 by ruling out the most adventurous proposals in that region, which require experimental or administrative data that does not exist in real estate. Psych method transfer (E) shows the same pattern at a lower level (HHI 0.112/0.119), confirming that the concentration is not driven by the specific psych paradigms selected but by a systematic gravitational pull toward the behavioural subfield.

Table 1: Research idea quality and spatial distribution by generation condition

Condition	N	Atypic.	Clust.	HHI	Data n/a	Periph.	NN dist.	Pred. cit.
A: Naïve	100	0.726 ± 0.180	21/51	0.070	0%	3.16	2.57	11.9
— *constr.*	100	0.706 ± 0.246	28/51	0.069	—	3.34	2.40	12.3
B: Full REE corpus	99	0.707 ± 0.248	23/51	0.074	0%	2.75	2.48	11.9
— *constr.*	100	0.697 ± 0.225	22/51	0.094	—	2.65	2.44	12.3
C: Cluster-seeded	51	0.720 ± 0.206	31/51	0.043	0%	2.89	2.48	11.5
— *constr.*	51	0.694 ± 0.198	30/51	0.045	—	2.58	2.37	11.8
D: Method trans. (econ/fin)	100	0.735 ± 0.147	25/51	0.089	0%	3.70	2.19	13.4
— *constr.*	100	0.718 ± 0.202	22/51	0.076	—	3.33	2.13	13.3
E: Method trans. (psych)	100	0.693 ± 0.210	21/51	0.112	13%	1.85	2.70	11.2
— *constr.*	99	0.656 ± 0.258	16/51	0.119	—	2.06	2.45	11.9
F: Paradigm trans. (econ/fin)	100	0.735 ± 0.158	23/51	0.080	0%	4.36	2.21	13.1
— *constr.*	100	0.763 ± 0.163	26/51	0.085	—	4.11	2.31	12.8
G: Paradigm trans. (psych)	100	0.632 ± 0.251	15/51	0.161	8%	1.74	2.67	11.2
— *constr.*	100	0.661 ± 0.243	20/51	0.100	—	1.90	2.43	11.8
H: Citation-guided	100	0.729 ± 0.179	27/51	0.066	0%	2.80	2.36	12.5
— *constr.*	99	0.687 ± 0.244	29/51	0.063	—	2.67	2.38	12.1
REE corpus (baseline)	*772*	—	*38/51*	*0.091*	—	*1.00*	*1.00*	—

Notes: Var. = a (unconstrained) or b (constrained; prompt requires the model to confirm data availability before proposing a design). Atypicality = mean ± SD; Shannon entropy of the 20-nearest-neighbour cluster distribution, normalised to [0, 1]. Cluster coverage = number of the 51 named HDBSCAN clusters receiving at least one assigned idea; noise papers (cluster −1) are included as a 52nd category in the HHI calculation. HHI = Herfindahl–Hirschman index of cluster shares; lower values indicate ideas spread more evenly across the research space. Data n/a = share of unconstrained ideas whose proposed data was rated non-existent or requiring primary collection (score 3 on a 1–3 scale); shown for unconstrained runs only. Periph. / REE = within-cluster peripherality normalised so REE corpus = 1.00; values above 1 indicate ideas land farther from cluster centres than typical corpus papers. NN dist. / REE = mean cosine nearest-neighbour distance in the sentence embedding space, normalised to the REE corpus mean; values above 1 indicate ideas are more distant from existing work than the average REE paper. Pred. cit. = mean predicted 5-year citation count assuming publication in REE, from OLS model trained on 1,074 papers across all five journals (REE baseline, age = 5); bold indicates the two highest values.

Figure 4: Generated Ideas in the Research Space — All 16 Generation Runs (part 1 of 2)

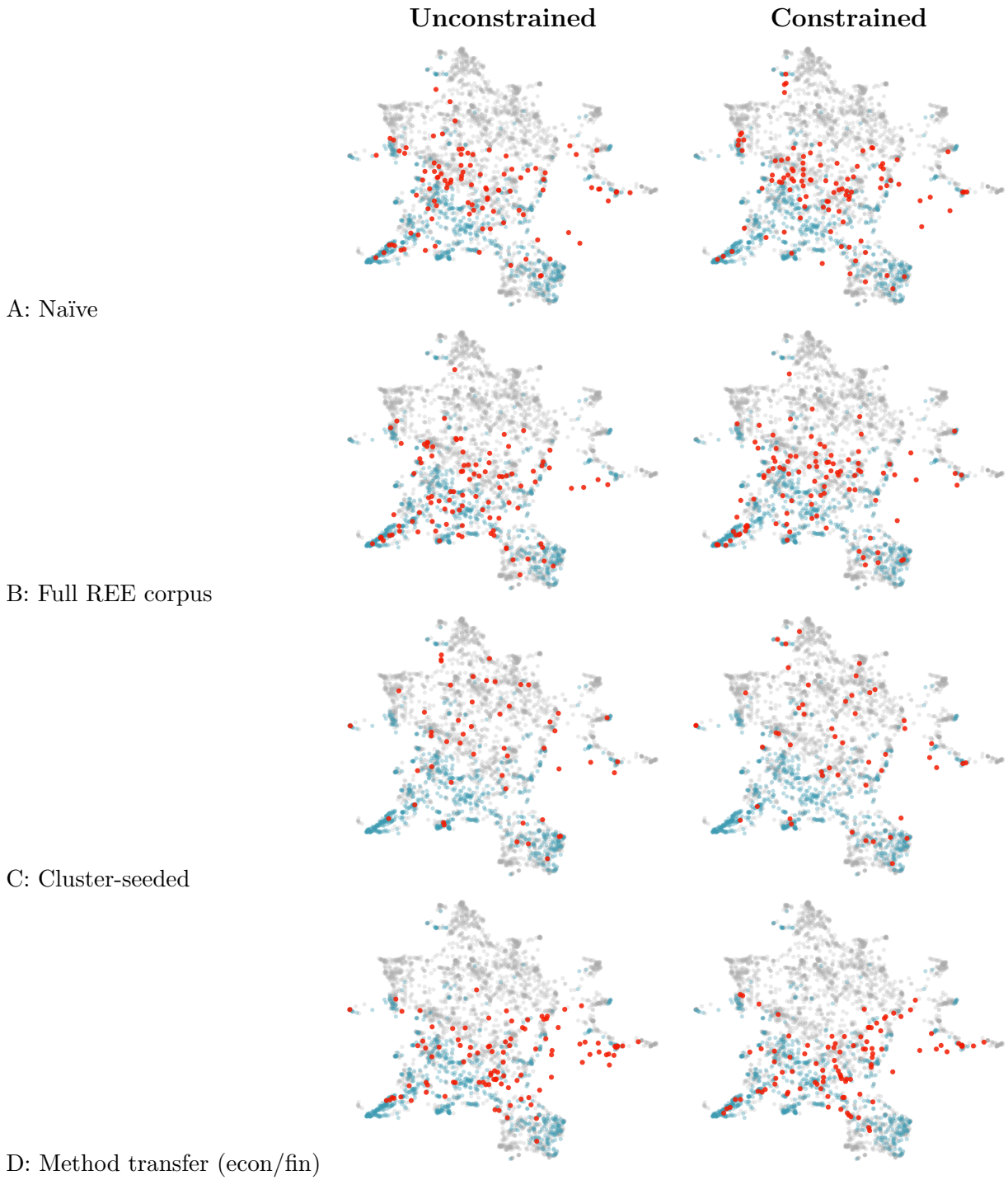
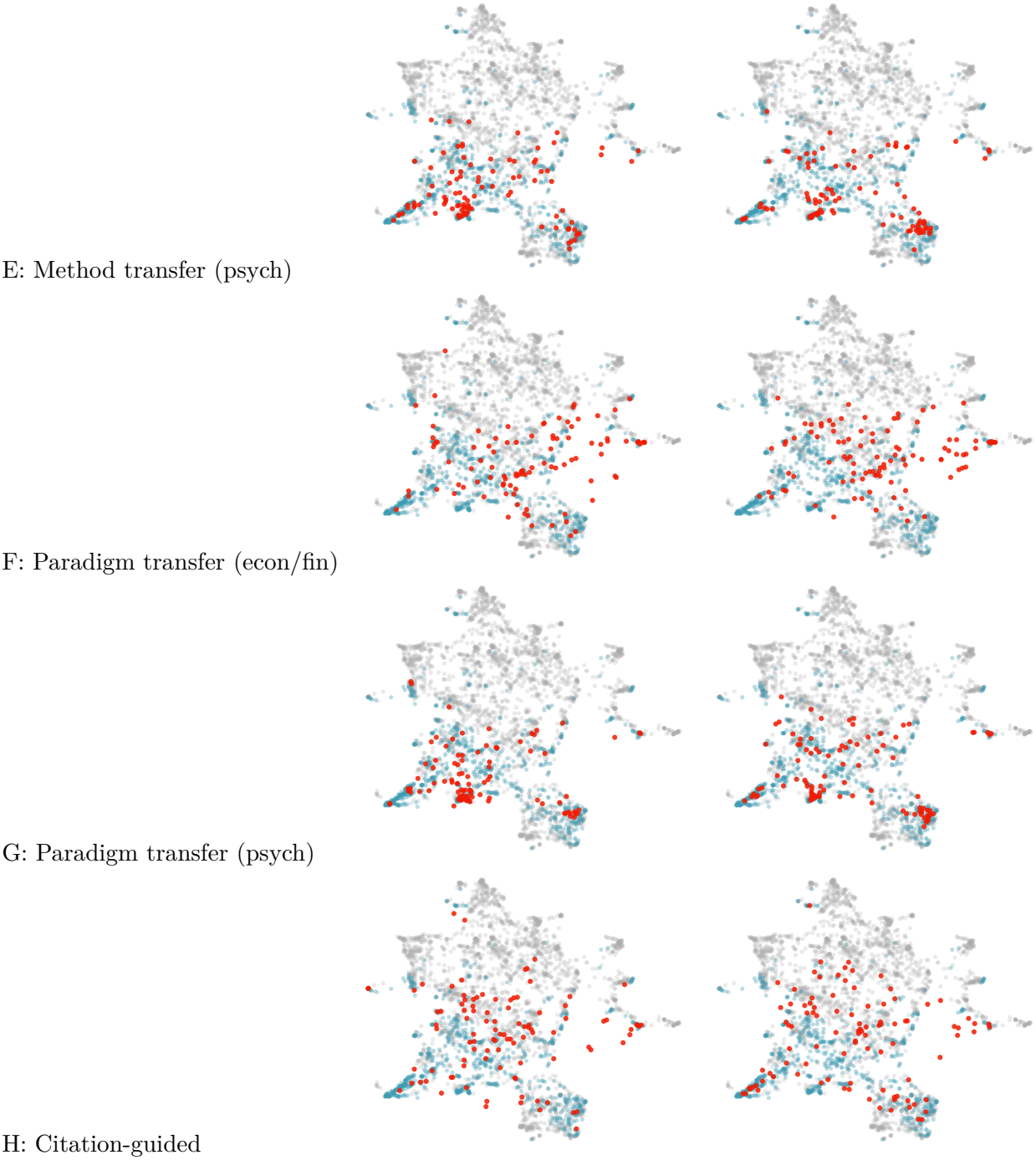


Figure 4 (continued)



E: Method transfer (psych)

F: Paradigm transfer (econ/fin)

G: Paradigm transfer (psych)

H: Citation-guided

Notes: Each panel projects generated ideas into the UMAP space fitted on the full corpus. Ideas are embedded using the same sentence transformer and projected via the pre-fitted UMAP model. Blue points mark the REE footprint; grey points mark the broader economics and finance space that generated ideas can reach but that was not part of the generation context. Unconstrained variants (a-panels) allow the model to generate freely; constrained variants (b-panels) include the instruction "only propose designs where you can confirm the required data exists and is accessible to an academic researcher." The five book-review entries in the far right of the corpus are excluded from the background. An interactive version of this figure — with zoom, full idea details, and a rating interface — is available at https://thies.github.io/idea-explorer/.

Econ/fin cross-field conditions (D and F) sit near or below the corpus baseline, suggesting that economic and financial methods, when applied to real estate questions, do not funnel toward any single subfield. The data constraint has little effect on their HHI, consistent with the interpretation that econ/fin ideas are grounded by design — the data they require is the same data real estate researchers already use.

Table 2 also reports data availability and within-cluster position. On data availability, each unconstrained idea's proposed data is classified as clearly accessible (score 1), uncertain (score 2), or non-existent (score 3), using a separate Haiku call rating the data description alone. Non-existent data is almost entirely a psych phenomenon: 13% of psych method-transfer ideas (E) and 8% of psych paradigm-transfer ideas (G) propose data that does not exist or would require primary collection, against 0% for every econ/fin-sourced and baseline condition. The psych-sourced conditions import methods whose execution in psychology relies on lab-based experimental infrastructure or purpose-built survey instruments that have no counterpart in real estate research. The model transplants the study logic successfully but fails to flag that the data collection apparatus that makes the original study executable is absent.

Within-cluster position compounds the story. Both spatial metrics in Table 2 are normalised to the REE corpus baseline (= 1.00). All generated ideas land substantially farther from their assigned cluster centroid than corpus papers do: relative peripherality ranges from 1.74× to 4.36× the REE benchmark. Ideas do not imitate cluster centres — they inhabit the edges. But the psych-sourced conditions are the least peripheral of all (E: 1.85×, G: 1.74× unconstrained), despite concentrating into fewer clusters. Concentration and centrality move together: the model picks a small behavioural territory and produces ideas that hug the existing core of that territory rather than pushing into unexplored sub-regions. Econ/fin paradigm transfer (F) is the opposite extreme — peripherality 4.36× the REE baseline — consistent with importing a complete study logic that lands in real estate clusters where it has no close precedent. Naïve generation (A: 3.16×) and citation-guided (H: 2.80×) fall in between: dispersed across many clusters but consistently peripheral within them. A parallel pattern is visible in published work. Papers from top finance and economics journals (JF, AER) that appear in the combined research space occupy fewer distinct clusters and are more concentrated than native REE papers: JF papers span 25 clusters with a Herfindahl index of 0.119, AER papers 29 clusters at 0.110, against REE's 37 clusters at 0.089. Cross-field imports in published work congregate in specific, established sub-regions of the research space, just as AI-generated paradigm-transfer proposals do — landing near where the methodology already has a foothold rather than opening genuinely new territory.

The nearest-neighbour distance in high-dimensional embedding space tells a different story from the within-cluster UMAP position, and the contrast is informative. In the sentence embedding space, all conditions are roughly 2.1–2.7× farther from their nearest existing paper than the typical REE paper is from its closest neighbour. Psych-sourced ideas (E: 2.70×, G: 2.67× unconstrained) have the largest high-dimensional gap despite being the most central in UMAP. This is not a contradiction: UMAP compresses the manifold to preserve neighbourhood structure at the cost of global distances, so points that appear near a cluster centre in the 2D projection can still occupy genuinely novel territory in the full embedding space. The implication is that psych-derived proposals are semantically distinct from existing real estate papers even when they land near the centroid of a behavioural cluster — their novelty is at the level of representation rather than topic placement. Econ/fin conditions (D: 2.19×, F: 2.21×) are the closest to existing work in high-dimensional space, which is consistent with their lower data non-existence rates: they propose studies that look like existing real estate research, using data that already exists, but push into unexplored corners of the within-cluster distribution.

The nearest-neighbour paper for each idea offers a partial window into what the model is drawing on. For naïve ideas, 36% land nearest to a paper in the top quartile of the citation distribution (REE corpus p75 = 43 citations), compared with 25% expected under a uniform draw — a modest but detectable citation bias. Providing the full REE corpus as context (condition B) barely shifts this: 35% of B's ideas have top-quartile nearest neighbours, nearly identical to A. The model's gravitational centres are not dislodged by context. Across all unconstrained conditions, the ten papers most frequently appearing as the nearest neighbour to a generated idea are concentrated around a handful of active research themes: climate and flood risk capitalisation, institutional landlord market power, remote-work effects on urban property values, appraisal bias, and information asymmetry in transactions. These are topics that have attracted both high citation counts and sustained media attention over the past decade — the model's implicit research agenda tracks the field's most visible recent frontier rather than its historical canon. Econ/fin cross-field conditions (D and F) land nearest to more highly-cited papers (median NN citations: 45 and 41 respectively, vs. 27 for naïve), which suggests that importing economic and financial methods does not open new terrain from scratch but applies them to questions the field already considers important.

Proximity to a high-citation nearest neighbour does not improve atypicality. The Spearman correlation between the citation count of an idea's nearest neighbour and its atypicality score is 0.07 ($p = 0.47$) for the naïve condition, and similarly flat across all conditions. The model generates near famous papers because those represent well-trodden

territory that is easy to extrapolate from, not because proximity to canonical work produces more atypical proposals.

The model's self-assessment of its own novelty is nearly uninformative. Of the 1,499 ideas generated across all sixteen runs, 89% receive a claimed novelty score of 3 out of 4, 10.5% score 2, and fewer than 1% score 4 — the effective variance in self-reported novelty is close to zero. The Spearman correlation between claimed novelty score and actual atypicality is 0.009 ($p = 0.74$): no detectable relationship with the measure that has demonstrated predictive validity for citations. More context does not produce better-calibrated self-assessment.

A one-way ANOVA on atypicality across the eight conditions is significant ($F = 2.97$, $p = 0.004$). The main separation is between the psych-sourced conditions and everything else. Psych paradigm transfer (G, mean atypicality 0.632) and psych method transfer (E, 0.693) trail the econ/fin and baseline conditions, which cluster between 0.707 and 0.735. Psychology methods and phenomena, transplanted to real estate, tend to land in a narrower region of the research space — likely because behavioural real estate already exists as a subfield, so psych-derived RE ideas settle near established work rather than at the frontier.

Among the non-psych conditions, differences are small. Econ/fin method transfer (D, 0.735) and paradigm transfer (F, 0.735) tie for the highest mean atypicality, followed by citation-guided (H, 0.729), naïve (A, 0.726), cluster-seeded (C, 0.720), and full-corpus (B, 0.707). The cluster scaffold shapes where ideas land but does not raise their atypicality: condition B spans 39 of 51 clusters versus 22 for naïve, but sits at the bottom of the non-psych atypicality distribution.

The distinction between conditions E and G is informative. Transplanting the whole study design (G) scores lower on atypicality than importing only the method (E), despite both drawing on psychology. Copying a complete psychological paradigm into real estate reproduces familiar behavioural research — loss aversion in selling, anchoring in appraisals — rather than opening new territory. Importing only the method and generating a fresh RE question produces more atypical combinations, even from the same source discipline.

Three findings run through the results. First, embedding-space scaffolding matters when it provides *methodological* rather than *topical* structure: telling the model what kinds of ideas already exist does not raise atypicality; telling it which empirical tools are underdeployed, and asking it to apply them, does. Second, the model's training priors are sticky — providing the entire REE corpus as context does not shift which existing papers generated ideas resemble, only how broadly they are distributed across the research space. Third, the model cannot reliably assess its own novelty: self-reported novelty scores

are nearly constant across all conditions and carry no information about where an idea actually lands in the embedding space. Any evaluation that relies on self-assessment is measuring confidence, not originality.

The A vs. B comparison also speaks to a concern raised about general-purpose scientific AI: that models trained predominantly on open-access literature may have inadequate knowledge of paywalled fields (Adam 2026; Gottweis et al. 2025). *Real Estate Economics* is published by Wiley and mostly paywalled; what a general-purpose LLM sees of it during training is mainly abstracts, working-paper versions, and indirect citations from top economics and finance journals that happen to reference the field. Yet condition B — which provides 786 full-text structured summaries explicitly — does not outperform condition A on atypicality, and the two conditions resemble the same existing papers. Two interpretations are consistent with this result. The first is that training data is adequate: the model has accumulated sufficient knowledge of REE's major themes through indirect exposure, so the full structured corpus adds breadth but not new ideas. The second is that training data is beside the point: generating atypical ideas requires an external scaffold to escape the field's gravitational pull, not better coverage of what is already inside it — and more domain knowledge, whether from training or context, only reinforces what the model already considers the productive territory. The cross-field conditions (D, F) succeed not by knowing REE better but by anchoring to external methods and forcing a departure. Real estate economics is also more heavily cross-cited with mainstream economics and finance than a more siloed field would be, so the model's indirect exposure to REE content may be unusually complete relative to a narrower specialty. Whether training-data gaps would bind more tightly in a genuinely isolated field remains an open question.

Predicted citation impact. As a further benchmark, expected five-year citations for each generated idea are estimated using a regression model calibrated on the full multi-journal corpus. The choice of measure here is constrained by the nature of the object being evaluated. All ex-post impact measures — raw citation counts, PageRank-style network centrality, the disruption index of Wu, Wang and Evans (2019) — require a published paper's realised citation network: which papers cited it, and whether those papers also cited its references. Generated ideas have no publication record and no citing papers. The disruption index in particular cannot be applied: it measures whether a paper's citers abandon its predecessors (disruption) or consolidate around them (conventionality), which requires both a declared reference list and forward citations — neither of which exists for a proposal that has never been submitted. A generated idea has no explicit pedigree; its nearest neighbours in embedding space reflect similarity, not intellectual debt, and cannot substitute for a reference list in the network calculation. The only tractable ex-ante

approach is a prediction model trained on realised outcomes for comparable published papers, which is what is estimated here. Atypicality serves as the embedding-space analogue of disruption: a proposal whose nearest neighbours span many clusters is more likely to bridge and displace existing strands than one that deepens a single one — the same intuition as the disruption index, derived from geometry rather than a realised citation network. The training set contains 1,074 papers from all five journals published between 2001 and 2021. The dependent variable is log(1 + citations); the specification is OLS with HC3-robust standard errors. The geometric predictors — cosine nearest-neighbour distance (nn_dist) and log within-cluster UMAP peripherality — are computed against the historically available corpus at each paper's publication year, avoiding look-ahead bias. The model includes log(age), atypicality, nn_dist, log(frontier_dist), a cluster-level mean log-citation term calibrated on REE papers (so predictions are anchored to REE citation norms), and journal fixed effects. All idea predictions are evaluated at the REE baseline (journal fixed effects set to zero) and at age = 5. $R^2 = 0.182$, in-sample Spearman r = 0.472.

The model estimated is:

$$\begin{aligned}\log(1 + c_i) = \alpha + \beta_1 \log(\text{age}_i) + \beta_2 \text{atypicality}_i + \beta_3 \text{nn\_dist}_i \\ + \beta_4 \log(\text{frontier\_dist}_i) + \beta_5 \bar{c}_{k(i)} + \sum_j \gamma_j \mathbf{1}[\text{journal}_i = j] + \varepsilon_i\end{aligned} \tag{1}$$

where c_i is the realised citation count, age_i is years since publication, atypicality_i is the Shannon entropy of the paper's 20-nearest-neighbour cluster distribution (a measure of how broadly the paper bridges research communities), nn_dist_i is the cosine distance to the nearest existing paper at publication time, $\log(\text{frontier\_dist}_i)$ is log within-cluster UMAP peripherality, $\bar{c}_{k(i)}$ is the mean log-citation for REE papers in the same cluster k (anchoring predictions to REE citation norms), and journal fixed effects capture outlet-level citation premia with REE as the baseline. All geometric predictors are computed against the historically available corpus at each paper's publication year, avoiding look-ahead bias.

Table 3 reports summary statistics for the 1,074 training papers; Table 4 reports the OLS estimates. The dominant predictors are log(age) ($\beta = 0.79$, $p < 0.001$), nn_dist ($\beta = -2.43$, $p < 0.001$), and the journal fixed effects (AER: +1.43, JF: +1.32, RFS: +1.18, JUE: +0.42, all $p < 0.001$ relative to REE). Atypicality is not significant ($\beta = 0.09$, $p = 0.50$). The cluster mean log-citation term captures what kind of work tends to attract citations in each part of the research space; once that is controlled, atypicality's marginal

Table 2: Summary statistics — citation prediction training set (N = 1,074)

Variable	Mean	SD	Min	Median	Max
log(1 + citations)	3.73	1.31	0.00	3.80	7.82
Citations (raw)	88.1	147.4	0	44	2,484
Atypicality	0.643	0.274	0.000	0.734	0.993
NN distance	0.219	0.071	0.028	0.212	0.497
log(age)	2.572	0.484	1.609	2.674	3.219
log(frontier dist)	−1.256	0.944	−4.310	−1.193	1.673
Cluster mean log-cit	3.220	0.652	0.693	3.209	5.656

Notes: Training set for OLS citation prediction model. 1,074 papers from REE (450), JUE (471), AER (64), RFS (64), and JF (25) published 2001–2021 with valid embeddings, cluster assignment, and citation count. Age measured as of 2026. Geometric features (atypicality, NN distance, frontier distance) computed against the historically available corpus at each paper's publication year.

Table 3: OLS citation prediction model

	Coefficient
log(age)	0.795***
	(0.074)
Atypicality	0.091
	(0.135)
NN distance	−2.433***
	(0.594)
log(frontier dist)	0.039
	(0.040)
Cluster mean log-cit	0.143**
	(0.070)
AER	1.433***
	(0.214)
JF	1.317***
	(0.279)
JUE	0.422***
	(0.080)
RFS	1.180***
	(0.187)
Constant	1.373***
	(0.365)
N	1,074
R^2	0.182

Notes: OLS with HC3-robust standard errors in parentheses. Reference category: REE. All journal dummies set to zero for idea predictions (REE baseline). Geometric predictors computed against the historically available corpus at each paper's publication year to avoid look-ahead bias. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

contribution is small. For scoring generated ideas the cluster placement is unknown ex ante — which cluster an idea eventually lands in is a downstream outcome — so atypicality remains the right ex-ante criterion.

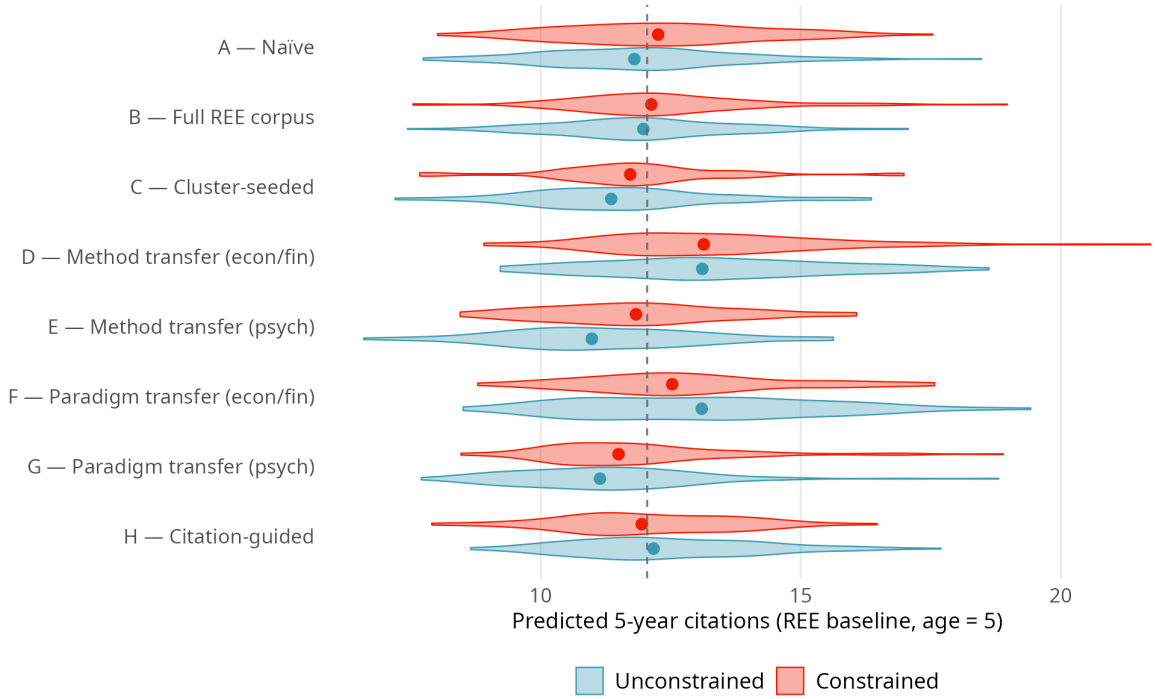
The predictions (Figure 5) show the econ/fin transfer advantage in citation terms. Method transfer from economics and finance (D) leads with a mean of 13.4 predicted five-year citations; paradigm transfer (F) follows at 13.1. Naïve (A: 12.3), full-corpus (B: 12.3), and citation-guided (H: 12.3) cluster together in the middle. Psych-sourced conditions (E: 11.5, G: 11.5) and cluster-seeded (C: 11.7) score lowest. The Spearman correlation between predicted citations and atypicality across all 1,499 ideas is -0.04 ($p = 0.12$): once the look-ahead bias in nn_dist is corrected, the two evaluation systems are independent.

The D advantage warrants decomposition. Because the model's dominant predictor is nn_dist ($\beta = -2.43$), any condition that produces ideas sitting closer to existing work will mechanically score higher on predicted citations — a proximity premium that is not obviously a measure of quality. Comparing D to naïve (A), the cluster-mean log-citation contribution is nearly identical across conditions (0.452–0.484), ruling out the possibility that D ideas land in intrinsically more citation-intensive clusters. Journal fixed effects do not enter here: all predictions are evaluated at the REE baseline. The difference is entirely in nn_dist: D unconstrained ideas have a mean cosine distance of 0.222 from their nearest existing paper, against 0.255 for A. Econ/fin method-transfer ideas are, by construction, closer matches to the existing literature — they adapt established empirical designs to real estate, producing proposals that resemble known work more than the unconstrained baseline does. The model rewards this proximity. Whether it is a genuine signal — established methods applied to new RE questions may indeed attract more citations from a field that values methodological rigour — or a reflection of the citation model's own training limitations is not fully resolvable from the data, but it is worth noting that the same proximity that flatters D's predicted citations is exactly what the atypicality criterion penalises.

5.5.3 Structured feasibility assessment

The two questions at the core of this evaluation — does the proposal address a gap in the literature, and is the proposed design executable? — are the same questions that automated peer review systems face for completed papers. Goyal et al. (2026) develop ScholarPeer, a multi-agent framework that tackles them through three specialised agents: one that situates a paper within the literature and assesses its contribution, one that evaluates methodological solidity, and a Q&A agent that probes specific claims. The

Figure 5: Predicted 5-Year Citations by Generation Condition



Notes: Violin plots with median dot for predicted 5-year citation count by generation condition and variant (unconstrained = teal; constrained = red). Predictions from OLS log(1 + citations) log(age) + nn_dist + log(frontier_dist) + atypicality + cluster_mean_log_cit + journal FE, estimated on 1,074 papers across all five journals (2001–2021); evaluated at the REE baseline (journal FE = 0) and age = 5. Geometric regressors (nn_dist, atypicality) computed against historically available papers at time of publication. Dashed vertical line marks the overall median.

approach here is the hypothetical analogue of their first two agents, applied to proposals rather than finished papers. ScholarPeer can ground its judgements in actual results; here the assessment is necessarily forward-looking.

To verify that the generation-stage data-existence constraint is doing its job, a structured external feasibility assessment is run in a separate call. Each idea is scored on four components (each 0–3): data precedent — has this type of data been used in published RE research?; data existence — does the proposed data actually exist and could a researcher access it?; method–data fit — is the proposed method structurally appropriate for the described data?; question plausibility — does the research question have a plausible economic mechanism? The composite structured score is the mean of the four components divided by 3.

The Pearson correlation between self-rated and structured feasibility is $r = 0.107$ ($p = 0.04$). The self-rated score at generation time carries almost no signal about actual feasibility — the model's confidence in its own proposals is uninformative.

Under the structured assessment, the main separation is between the non-psychology conditions (A, B, C, D, F, H, averaging 0.471–0.483) and the psychology conditions (E: 0.397, G: 0.370). The ANOVA is significant ($F = 8.28$, $p < 0.0001$). Psych-sourced conditions score lowest because they import methods whose execution relies on lab-based infrastructure or purpose-built survey instruments that have no counterpart in real estate research.

The structured components reveal where ideas fail. Data existence is the universal bottleneck: it scores 1.4–1.9 out of 3 across all conditions, well below question plausibility (2.2–2.8). Ideas are generally coherent — the model constructs plausible economic mechanisms — but frequently propose data that does not exist or is inaccessible. Method-transfer ideas score worst on method–data fit (1.52 vs. 1.74–1.90 for other conditions): the econometric designs imported from finance and economics often require data or identifying variation that real estate settings cannot provide. The rater's written reasoning on the lowest-scoring ideas is pointed — ideas requiring AVM deployment timelines from patent filings, cross-domain linked administrative records, or private-platform origination data that are not systematically accessible. Psych method-transfer (D) scores lowest on data precedent (1.10): the imported methods are genuinely novel in the RE literature, but novelty of method does not imply the data needed to execute them is available.

Two econ/fin method-transfer ideas score at the floor (feasibility 0.5) despite clear intellectual merit. The first asks whether randomly providing homeowners with expert long-run price forecasts shifts home improvement spending and portfolio rebalancing — a clean causal design, the right question, and an estimable treatment effect. The execution

requires recruiting 5,000–10,000 homeowners through a mortgage servicer or housing counselling non-profit and tracking them via building permit records, credit files, and brokerage accounts for two years. No such experiment exists. The second proposes to test whether employer-sponsored down payment assistance raises homeownership and lowers worker turnover, exploiting variation across firms; the data require an HR analytics provider to link employee records to credit bureau homeownership data and firm productivity metrics. Again, the question is good and the identification strategy is sensible. Both are publishable ideas. Neither is executable from a university office. The model gets the science right and the logistics wrong.

The constrained prompt does not solve the problem. The data-existence instruction reduces data non-existence rates for psych conditions (G drops from 8% to 0% flagged) but the structured feasibility scores barely move: psych conditions score 0.370–0.397 under the structured assessment regardless of whether the constraint was active. The constraint changes what the model *says* about its data requirements, not what it actually proposes. A post-hoc structured assessment remains the only reliable filter.

All of this is still rough. Atypicality, predicted citations, and structured feasibility scores are proxies, and the proof is in the pudding: what matters is whether domain experts find the ideas original, executable, and worth pursuing. To collect that evidence, the 1,499 generated ideas are made available in an interactive web application at https://thies.github.io/idea-explorer/, where researchers can browse the full embedding map, read each proposal in detail, and rate it on originality, feasibility, and relevance. Once a sufficient number of expert ratings has been collected, the human scores can serve as a second evaluation layer — one grounded in domain knowledge and research experience rather than bibliometric structure. The correlation between the automated scores and human judgements will itself be a measure of how much signal the embedding-space proxies actually carry.

Whether that correlation will be high enough to validate the approach is not obvious. In the natural sciences, evaluation is ultimately grounded in experimental outcomes: a drug candidate either works or it does not. Social science and economics lack that bench test, and real estate economics — straddling institutional economics, urban economics, and finance — spans enough methodological territory that expert opinion will vary. Reaching the level of evaluative certainty available in other fields will remain a challenge.

6 Conclusion

Excess-word analysis on roughly 101,000 real-estate and housing papers indexed by OpenAlex shows the same sharp 2023 inflection that Kobak et al. (2025) documented for biomedical literature. "Delve" ran at 0.006 occurrences per 1,000 abstract words across 2010–2022 and reached 0.127 in 2024 — a twenty-fold increase. The rate of all LLM-indicator words combined tripled in two years. Real estate scholars adopted AI-assisted writing at the same pace as biomedical researchers.

This paper constructs the first full-text, content-based map of the real estate finance and economics literature, characterising each paper along six structured dimensions — research question, method, data, sector, finding, and AI role — extracted by LLM from full text. The result is not a conventional review but measurement infrastructure: a coordinate system against which new proposals can be located.

Identifiable AI/ML papers are absent from *Real Estate Economics* before 2020 and reach 5–9 per cent of annual output only from 2023 onward. This is a lower bound: it counts only papers where AI performs an otherwise infeasible task and where the method is explicit. A larger grey zone — AI used for data collection, literature screening, transcription, or other traditional support tasks that leave no trace in the methods section — is not captured by keyword detection and not reflected in these shares. Of the identifiable AI papers, nearly half appear in a single cluster (hedonic pricing), and all use AI as a measurement or prediction tool. None applies it for causal inference support or studies AI as an economic phenomenon, the two roles that have already diffused into neighbouring economics and finance journals. On what is observable, REE is a late and selective adopter.

Generated ideas are evaluated on atypicality — Shannon entropy of the cluster distribution among an idea's 20 nearest neighbours, measuring how broadly a proposal bridges existing research communities rather than extending a single one. This operationalises the Uzzi et al. (2013) argument that high-impact science combines conventional foundations with atypical elements. The citation prediction model confirms that the right dimension to maximise is not raw distance from the frontier: nn_dist is a significant negative predictor of citations ($\beta = -2.43$, $p < 0.001$), meaning proposals that sit far from all existing work attract fewer citations, not more. Atypicality is the ex-ante signal for community-bridging, computable before cluster placement is known. Feasibility is handled at the generation stage through a hard data-existence constraint in the prompt; a structured post-hoc assessment confirms self-rated feasibility carries almost no signal about actual executability ($r = 0.107$ with structured scores).

In total, 1,499 research ideas are generated under eight conditions and evaluated on atypicality. The main separation is between the psych-sourced conditions (E: 0.693, G: 0.632) and the econ/fin and baseline conditions (0.707–0.735); ANOVA $F = 2.97$, $p = 0.004$. A citation prediction model trained on 1,074 papers across all five journals shows econ/fin method transfer (D) leading on predicted five-year citations (13.4 vs. 11.5–12.3), but decomposition reveals this is driven entirely by lower nn_dist — D ideas sit closer to existing work by construction, and the citation model rewards proximity. Atypicality and predicted citations are independent (Spearman $r = -0.04$, $p = 0.12$): D ranks first on predicted citations precisely because it is least atypical. The model's own confidence about its proposals carries no information about actual executability; data existence is the binding constraint, and the generation-stage constraint changes what the model says about its data requirements more than what it actually proposes.

Set aside the measures and ask the simpler question: are the AI-generated proposals any good? Some are not. A non-trivial share land squarely on problems the field has already answered, rediscovering questions settled in papers published a decade or two ago. But that is also true of human research proposals — most grant applications and seminar pitches are not original either. The best ideas, particularly those generated through method transfer from economics and finance, score comparably on atypicality to the median published paper in REE. A handful are genuinely striking: original combinations with plausible mechanisms and accessible data that, to the best of the author's knowledge, have not been attempted. The boundary between research assistant and principal investigator is narrowing. Whether it disappears is a question the next generation of purpose-built scientific AI systems — rather than the general-purpose models used here — will answer.

All generation in this paper uses a general-purpose LLM (Claude Sonnet 3.5). Dedicated scientific AI systems are now in active deployment: Google's AI co-scientist, launched as a trusted tester program in February 2025, is a purpose-built multi-agent architecture that has already produced experimentally validated hypotheses in biomedical research (Gottweis et al. 2025; Adam 2026). How much of this is a property of language models as a class, and how much is a transient feature of the specific general-purpose systems available during 2024–2025, is not answerable from this data. The ideation-execution gap documented by Si, Hashimoto and Yang (2025b) — AI ideas look better before implementation than after — may narrow or widen as purpose-built systems replace general-purpose ones. The field-mapping infrastructure and the atypicality-as-criterion approach are not model-specific; the generation results are.

There is a near-universal pattern in how AI tools get adopted: if a system appears to offer a useful shortcut for a task it was not built for, people use it anyway, ignoring the

inconvenient truth that they are in unchartered territory. It takes substantial effort to do even basic due diligence. In our case, researchers are already using general-purpose LLMs to generate research ideas. That will not stop. The more productive question — the one this paper attempts to answer — is what those systems actually produce when they are asked to guide research projects, under what conditions the output is worth anything, and where the failure modes lie.

7 Data and code availability

All code will be made available at the project repository. Structured extraction outputs, embeddings, and cluster assignments can be shared subject to publisher licensing constraints. Raw PDFs cannot be redistributed due to copyright limitations.

8 Acknowledgements

I am grateful to Lucy McGregor and Colin Lizieri for the clarity, insight, and constructive criticism they brought to discussions of the role of AI in teaching and research.

References

Adam, D, "The AI co-scientist is here," *Nature Medicine*, 2026, *32*, 772–775.

Angrist, J. D. and J.-S Pischke, *Mostly Harmless Econometrics*, Princeton University Press, 2009.

Athey, S. and G. W Imbens, "Journal of Economic Perspectives," 2017.

_ and _ , "Annual Review of Economics," 2019.

Baek, J., "ResearchAgent: Iterative Research Idea Generation over Scientific Literature," 2025.

Bahoo, S., J. Cuñado, and K Gupta, "Artificial Intelligence in Economics Research: What Have We Learned? What Do We Need to Learn?," *Journal of Economic Surveys*, 2025.

Bartik, A. W., A. Gupta, and D Milo, "The Costs of Housing Regulation: Evidence from Generative Regulatory Measurement," 2025. Working paper.

Boiko, D. A., R. MacKnight, B. Kline, and G Gomes, "Nature," 2023.

Calainho, F. D., A. M. van de Minne, and M. K Francke, "A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate," *Journal of Real Estate Finance and Economics*, 2024, *68*, 624–653.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J Robins, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 2018, *21* (1), C1–C68.

Cissé, A., M. E. Cooper, M. Zhu, X. Evangelopoulos, and A. I Cooper, "Can We Automate Scientific Reasoning in Closed-Loop Experiments using Large Language Models?," 2026.

Donthu, N., S. Kumar, D. Mukherjee, N. Pandey, and W. M Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, 2021, *133*, 285–296.

Fama, E. F., L. Fisher, M. C. Jensen, and R Roll, "The Adjustment of Stock Prices to New Information," *International Economic Review*, 1969, *10* (1), 1–21.

Fortunato, S., C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, and A.-L. Barabási, "Science of science," *Science*, 2018, *359* (6379), eaao0185.

Foster, J. G., A. Rzhetsky, and J. A Evans, "Tradition and Innovation in Scientists' Research Strategies," *American Sociological Review*, 2015, *80* (5), 875–908.

Francke, M. and A van de Minne, "Combining machine learning and econometrics: Application to commercial real estate prices," *Real Estate Economics*, 2024, *52*, 1308–1339.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *Journal of Finance*, 2022, *77* (1), 5–47.

Gentzkow, M., B. Kelly, and M Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gottweis, J., W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, and V ... Natarajan, "Towards an AI co-scientist," 2025.

Goyal, P., M. Parmar, Y. Song, H. Palangi, T. Pfister, and J Yoon, "ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review," 2026.

Gu, S., B. Kelly, and D Xiu, "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Imbens, G. and T Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 2008, *142* (2), 615–635.

Kelly, B. and D Xiu, "Financial Machine Learning," 2023. Working paper.

Kobak, D., R. González-Márquez, E.-Á. Horvát, and J Lause, "Delving into ChatGPT usage in academic writing through excess word analysis," *Science Advances*, 2025, *11* (3), eadt3813.

Lee, D. S. and T Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

Leow, K. and T Lindenthal, "Enhancing Real Estate Investment Trust Return Forecasts using Machine Learning," *Real Estate Economics*, 2025.

Lindenthal, T. and E. B Johnson, "Machine Learning, Architectural Styles and Property Values," *Journal of Real Estate Finance and Economics*, 2021.

Lopez, L. A., S. J. McCoy, and V Sah, "Steering consumers to lenders in residential real estate markets," *Real Estate Economics*, 2022, *50* (4), 1596–1641.

Lorenz, F., N. Kok, and co authors, "Interpretable Machine Learning for Real Estate Market Analysis," *Real Estate Economics*, 2023.

Lu, C., C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J Clune, "Towards end-to-end automation of AI research," *Nature*, 2026.

Merchant, A., S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D Cubuk, "Scaling deep learning for materials discovery," *Nature*, 2023, *624*, 80–85.

Naik, N., R. Raskar, and C. A Hidalgo, "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance," *American Economic Review*, 2016, *106* (5), 128–132.

Rosen, S, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Shen, L. and S. L Ross, "Information value of property description: A machine learning approach," *Journal of Urban Economics*, 2021, *121*, 103299.

Si, C., T. Hashimoto, and D Yang, "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers," in "ICLR 2025" 2025a.

_ , _ , and _ , "The Ideation–Execution Gap in AI-Generated Research Ideas," 2025b.

Su, H., R. Chen, S. Tang, X. Yin, J. Zheng, B. Qi, Q. Wu, H. Li, W. Ouyang, P. Torr, B. Zhou, and N Dong, "Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system," *ACL 2025*, 2025.

Tekouabou, S. C. K., Ş. C. Gherghina, E. D. Kameni, Y. Filali, and K Idrissi Gartoumi, "AI-Based Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey," *Archives of Computational Methods in Engineering*, 2023, *31*, 1079–1095.

Tshitoyan, V., J. Dagdelen, L. Weston, K. A. Persson, G. Ceder, and A Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, 2019, *571*, 95–98.

Uzzi, B., S. Mukherjee, M. Stringer, and B Jones, "Atypical Combinations and Scientific Impact," *Science*, 2013, *342* (6157), 468–472.

Wager, S. and S Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Wan, W. X. and T Lindenthal, "Testing Machine Learning Systems in Real Estate," *Real Estate Economics*, 2023.

Wang, J., R. Veugelers, and P Stephan, "Bias against novelty in science: A cautionary tale for users of bibliometric indicators," *Research Policy*, 2017, *46* (8), 1416–1436.

Wu, L., D. Wang, and J. A Evans, "Large teams develop and small teams disrupt science and technology," *Nature*, 2019, *566*, 378–382.

Yakura, H., E. Lopez-Lopez, L. Brinkmann, I. Serna, P. Gupta, I. Soraperra, and I Rahwan, "Empirical evidence of Large Language Model's influence on human spoken communication," 2025.