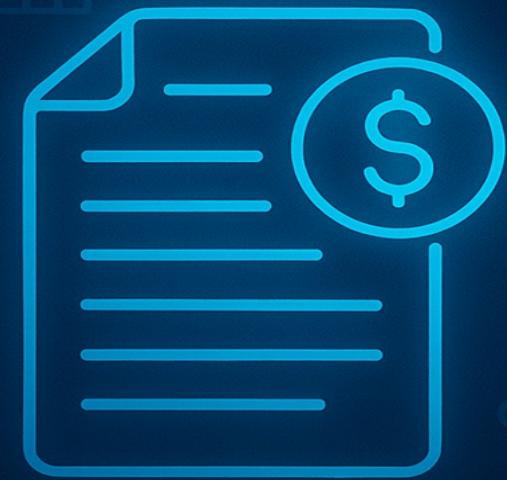


AI in Finance

Arpit Gupta (NYU Stern)
Spring 2026
Session 2

Financial Document Intelligence



Financial Document Intelligence — 1

- A key application of AI/LLMs in finance is in a topic area we can think of as **document intelligence**
- This is the task of turning a set of documents into actionable insights
- Starts with **extraction** and reconstruction of a document base
 - Ie, 10-Ks/10-Qs, earnings call, prospectuses, ratings reports, regulations
 - Keep track of tables, text, metadata associated with underlying document
- Then we do **semantic understanding** and normalization
 - Mapping phrases into stable concepts

Financial Document Intelligence — 2

- Next is **retrieval** and reasoning over this corpus
 - In practice, this typically means Retrieval-Augmented Generation (RAG), with appropriate document “chunking” and strategies to ensure accurate referencing
- The value from this is unlocked through **workflow automation**
 - Ie, what insights or actions can be triggered (covenant violation, flag fraud)
- Finally there needs to be a reliability and governance layer for financial applications
 - Requires **provenance** (where did this fact come from), evaluation, and controls for deployment

Customizing AI for Finance

Three Levels of AI Customization

Approach	Investment	Depth of Change	When to Use
Prompt Engineering	Minutes to Hours	No Model Change, guides behavior	Most Use Cases
Fine-tuning	Thousands of dollars	Embeds domain knowledge	Specialized tasks, consistent style
Pre-training	\$10m-\$100m+	Build from Scratch	Only major labs

Most Financial Applications Relate to Prompting or Fine-Tuning

- Foundational models contain broad knowledge across domains, including finance
- Customization is typically about activating, extending, and applying that knowledge
- Pre-training creates general capabilities (language, reasoning, world knowledge). Quality filtering here is essential.
- Fine-tuning and prompting specialize those capabilities for your specific use case

Reinforcement Learning

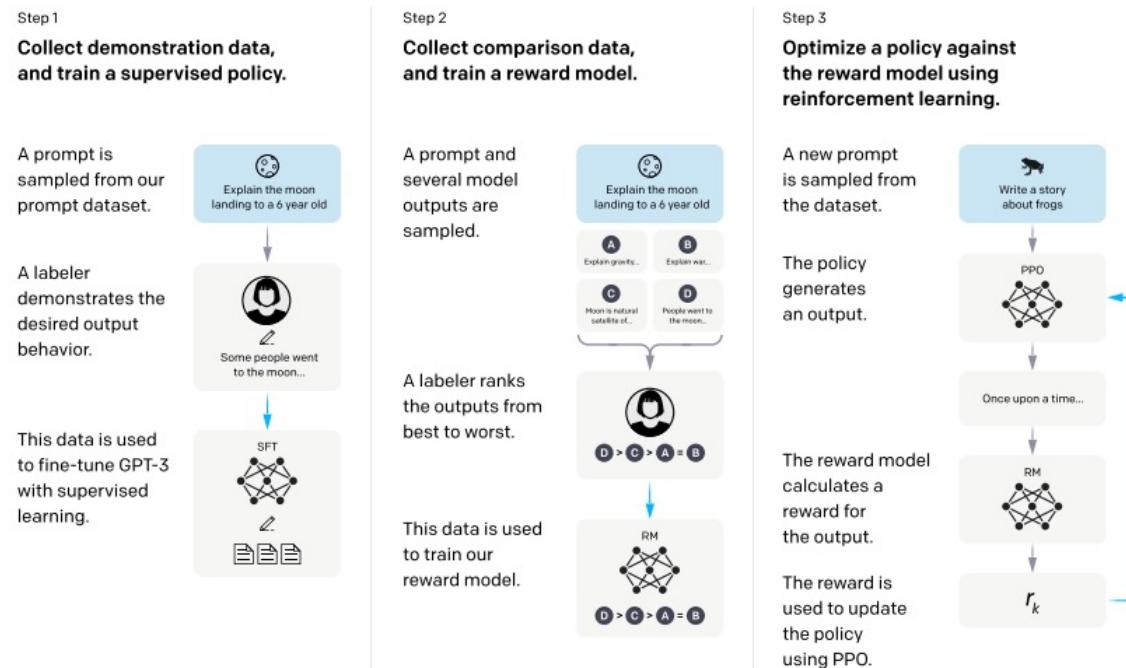


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

Reinforcement Learning

- An additional component of training to make the final output valuable by users is **reinforcement learning**
 - Supervised training (demonstrated output)
 - Reward model (human labeled comparisons)
 - Optimized Policy (systematic ranking of model outputs)
- Helps to produce answers closer to what people want in revealed preference

Prompt Engineering

- Effective prompts do three things:
 1. Set the context: what documents, data, or firms are relevant?
 2. Define the role: "you are a credit analyst" vs generic query
 3. Structure the reasoning: "Think step by step" (chain of thought) can improve reasoning
- Explicit role-setting matters less with newer models, but context remains essential

Prompting Example

Can ChatGPT Forecast Stock Price Movements?

Return Predictability and Large Language Models *

Alejandro Lopez-Lira and Yuehua Tang

University of Florida

First Version: April 6, 2023; This Version: October 28, 2025

Abstract

We document the capability of large language models (LLMs) like ChatGPT to predict stock market reactions from news headlines without direct financial training. Using post-knowledge-cutoff headlines, GPT-4 captures initial market responses, achieving approximately 90% portfolio-day hit rates for the non-tradable initial reaction. GPT-4 scores also significantly predict the subsequent drift, especially for small stocks and negative news. Forecasting ability generally increases with model size, suggesting that financial reasoning is an emerging capacity of complex LLMs. Strategy returns decline as LLM adoption rises, consistent with improved price efficiency. To rationalize these findings, we develop a theoretical model that incorporates LLM technology, information-processing capacity constraints, underreaction, and limits to arbitrage.

Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer “YES” if good news, “NO” if bad news, or “UNKNOWN” if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of _company_name_ in the short term?

Headline: _headline_

Jagged Frontier of AI Abilities

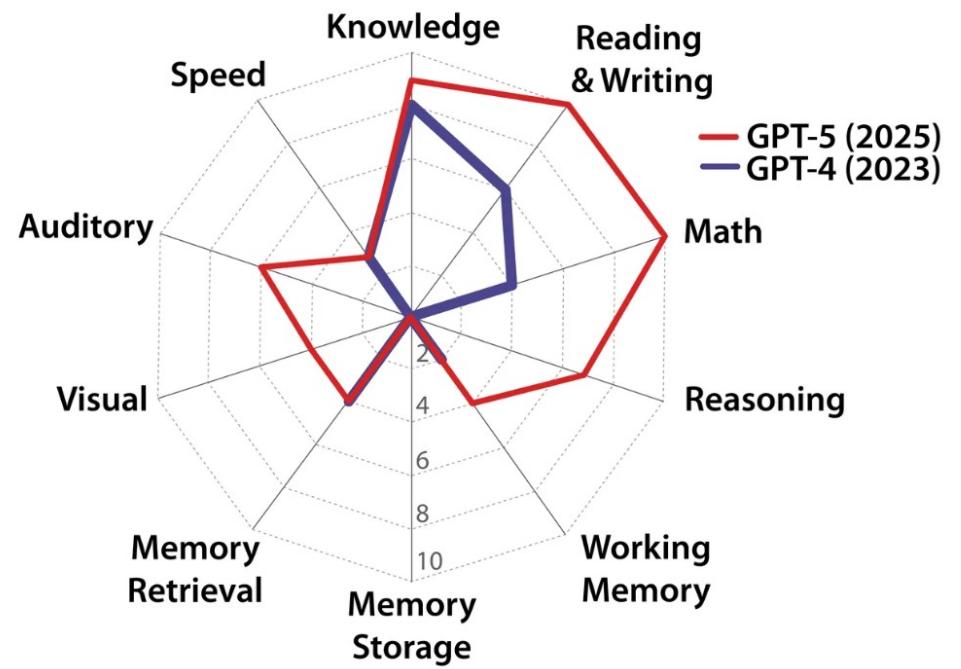
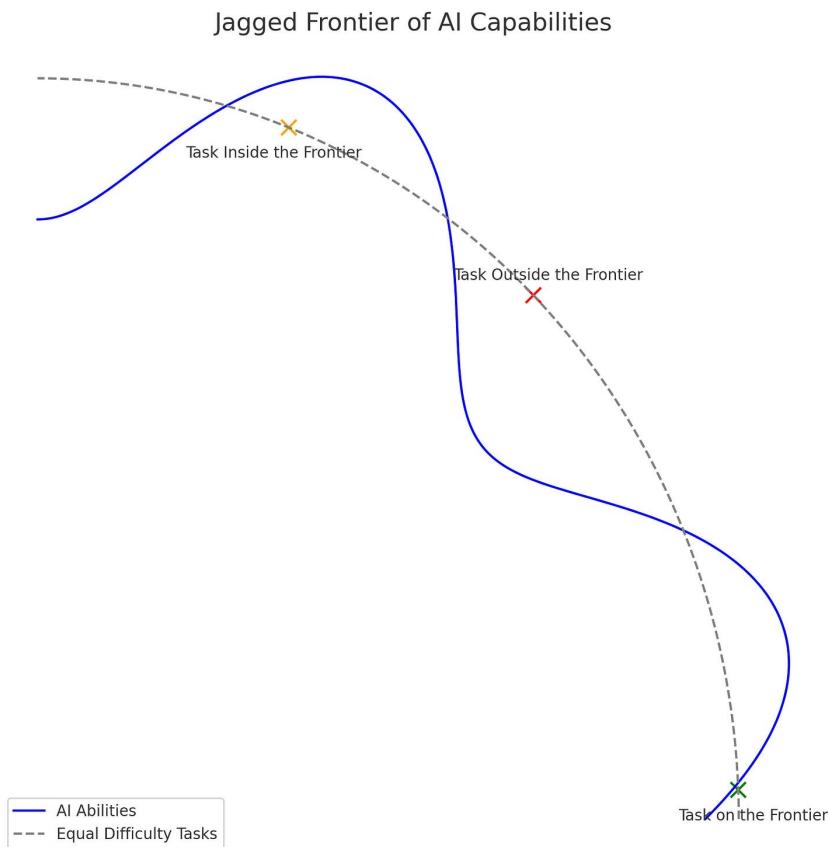
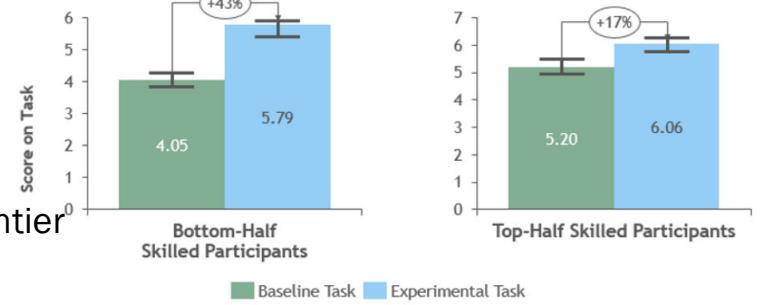
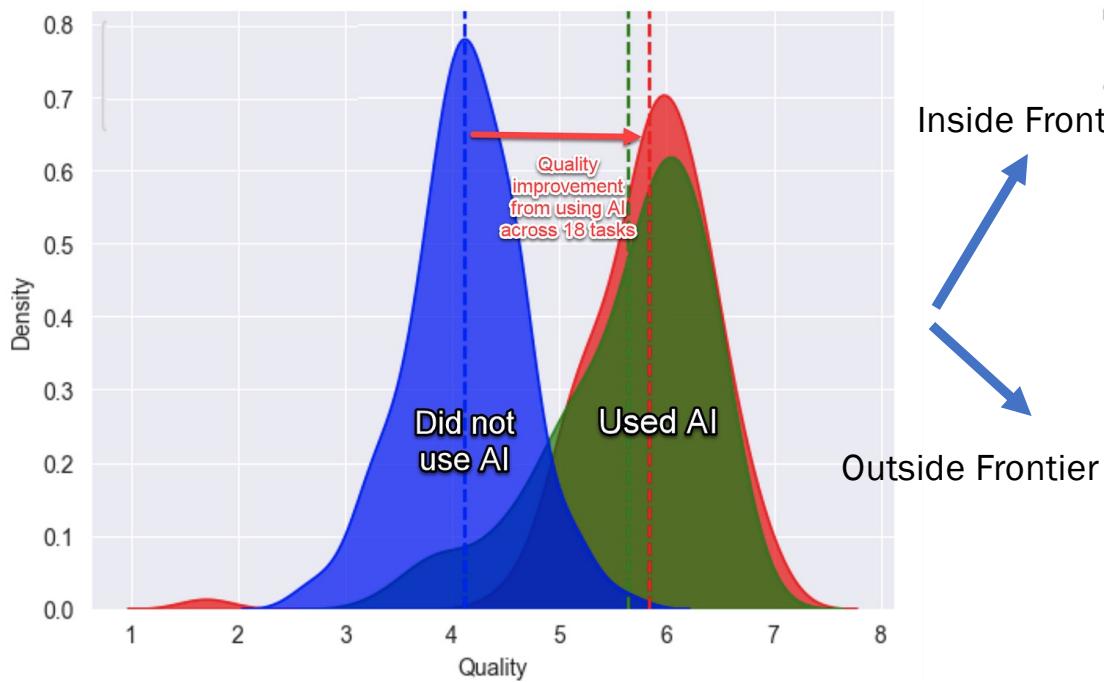


Figure 5: Bottom-Half Skills and Top-Half Skills - Inside the Frontier

Around Jagged Frontier

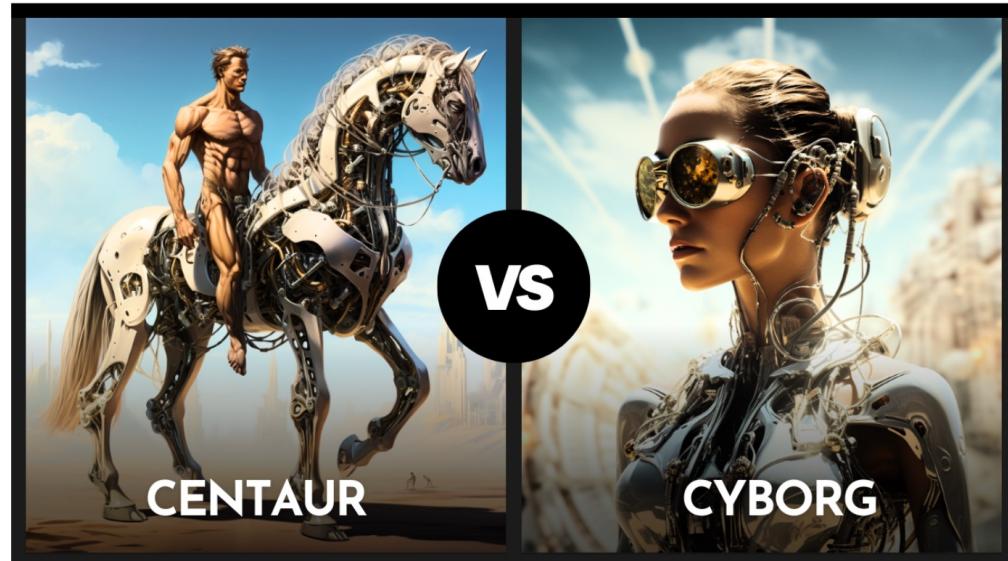


"Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality" by Dell'Acqua et al.

Notes: This figure displays average performance for the task outside the frontier. It reports the percentage of subjects in each experimental group providing a correct response in the experimental task.

Centaur vs Cyborg

- Centaur: division of labor between human and AI task
 - Clear context and prompt, derived from human-produced content, more important
- Cyborg: blending and integrating huma + machine tasks
 - Iterative and back and forth; precise prompt less relevant than the entire interaction



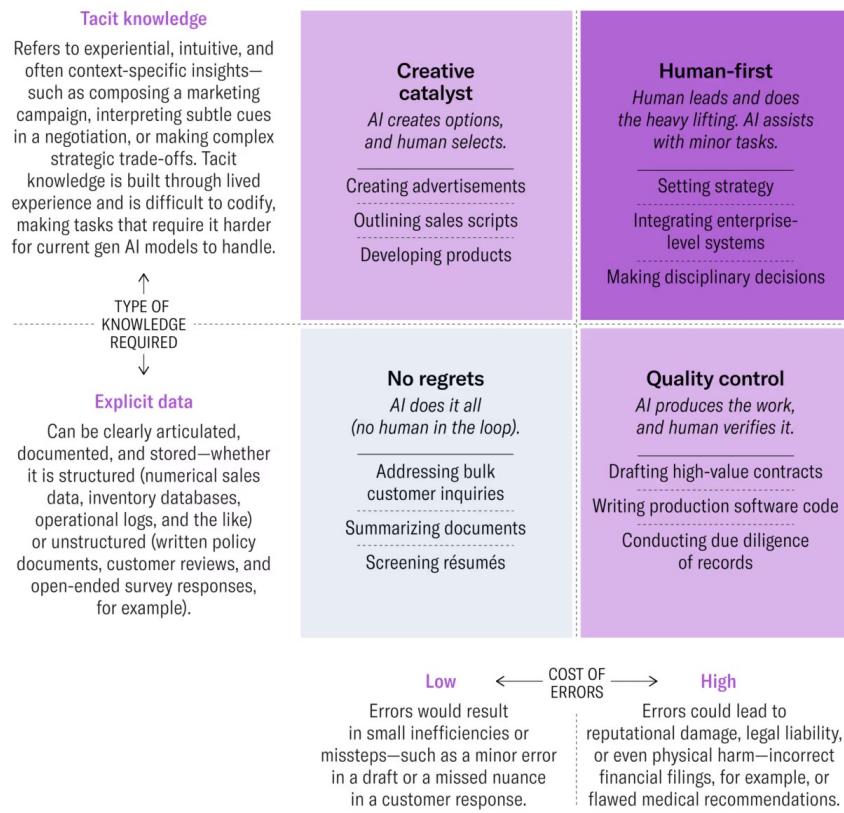
Broader Limitations to AI Adoption

1. Capability bottleneck
 - I.e., the jagged frontier
2. Process bottleneck
 - Institutional and regulatory constraints around AI adoption
3. Verification bottleneck
 - Humans necessary for final accuracy check

A Framework for Choosing Where and How to Use Gen AI

To determine how best to deploy gen AI in your organization, you must first identify the tasks being performed and evaluate them on two factors: What are the consequences if gen AI makes a mistake? What type of knowledge is needed? Once you locate each task within the matrix, you can use the appropriate level of gen AI support.

As the framework illustrates, some tasks are best handled by AI alone, others through human-AI collaboration, and some still require purely human judgment. The key is understanding which tasks remain distinctly human.



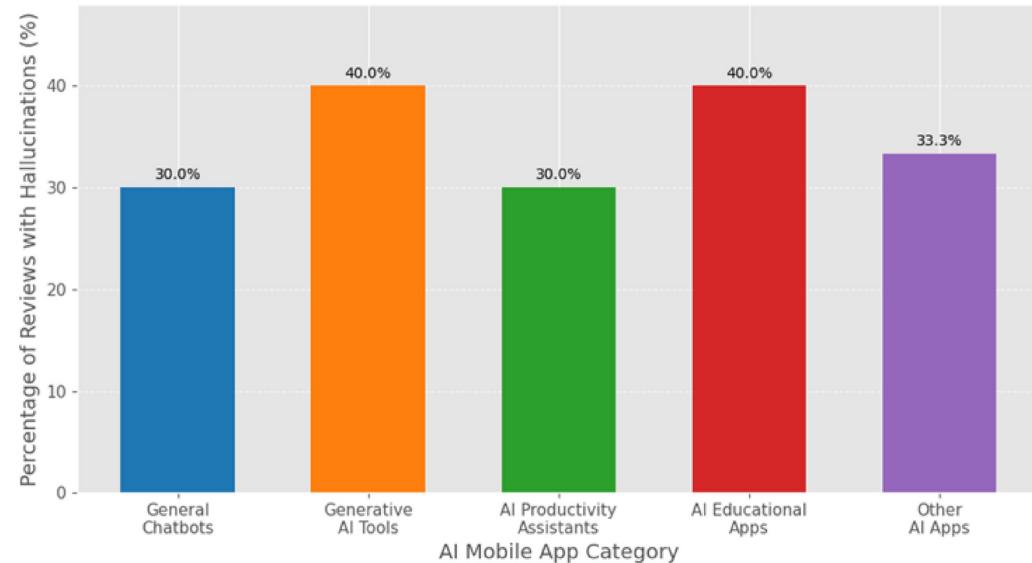
Source: Bharat N. Anand and Andy Wu

HBR

Truth and Fiction

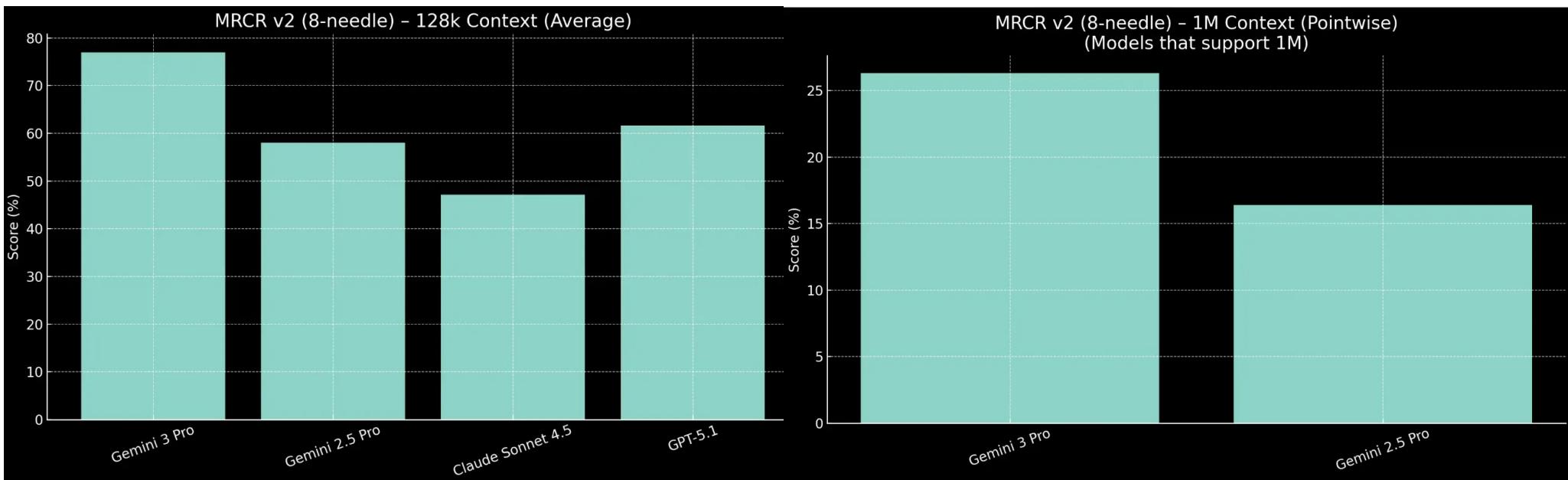
The Problem: Hallucination

- LLMs confidently state wrong information inherently, as a function of how they work
 - Outdated information drawn from training data
 - Never learned domain-specific information
 - Any time fiction is a more plausible next-token prediction than reality



"My AI is Lying to Me": User-reported hallucinations in AI mobile apps reviews

Bigger Problem in Longer Contexts



When is Hallucination less of a problem?

- When output is easy to verify (i.e., “write code to generate a SQL query.” Does the code work or not?)
 - Easier to do when you are a domain expert in the area
- Creative/narrative tasks
- Work that depends on the “best” ideas (i.e., brainstorming)
 - “give me 15 ways to...”
- Downside of errors is low
- When you are simulating possible responses
 - Ie, first pass at what a hostile or naïve reviewer/customer might think
- Data is used as part of statistical inference approach which tolerates errors

When is it a Bigger Problem?

- Ground truth is important to **make a decision** in the real world
 - Often the case in Finance applications!
- The human may fall asleep at the wheel; poor oversight ->

Falling Asleep at the Wheel: Human/AI
Collaboration in a Field Experiment on HR
Recruiters

Fabrizio Dell'Acqua

Laboratory for Innovation Science, Harvard Business School

Abstract

I investigate how firms should design human/AI collaboration to ensure human workers remain engaged in their activities. I developed a formal model that explores the tension between AI quality and human effort. As AI quality increases, humans have fewer incentives to exert effort and remain attentive, allowing the AI to substitute, rather than augment their performance. Thus, high-performing algorithms may do worse than lower-performing ones in maximizing combined output. I then test these predictions using a pre-registered field experiment where I hired 181 professional recruiters to review 44 resumes. I selected a random subset of screeners to receive algorithmic recommendations about job candidates, and randomly varied the quality of the AI predictions they received. I found that subjects with higher quality AI were less accurate in their assessments of job applications than subjects with lower quality AI. On average, recruiters receiving lower quality AI exerted more effort and spent more time evaluating the resumes, and were less likely to automatically select the AI-recommended candidate. The recruiters collaborating with low-quality AI learned to interact better with their assigned AI and improved their performance. Crucially, these effects were driven by more experienced recruiters. Overall, the results show that maximizing human/AI performance may require lower quality AI, depending on the effort, learning, and skillset of the humans involved.

Over time, AI use can Impair Skill Formation

How AI Impacts Skill Formation

Judy Hanwen Shen* Alex Tamkin†

January 29, 2026

Abstract

AI assistance produces significant productivity gains across professional domains, particularly for novice workers. Yet how this assistance affects the development of skills required to effectively supervise AI remains unclear. Novice workers who rely heavily on AI to complete unfamiliar tasks may compromise their own skill acquisition in the process. We conduct randomized experiments to study how developers gained mastery of a new asynchronous programming library with and without the assistance of AI. We find that AI use impairs conceptual understanding, code reading, and debugging abilities, without delivering significant efficiency gains on average. Participants who fully delegated coding tasks showed some productivity improvements, but at the cost of learning the library. We identify six distinct AI interaction patterns, three of which involve cognitive engagement and preserve learning outcomes even when participants receive AI assistance. Our findings suggest that AI-enhanced productivity is not a shortcut to competence and AI assistance should be carefully adopted into workflows to preserve skill formation – particularly in safety-critical domains.

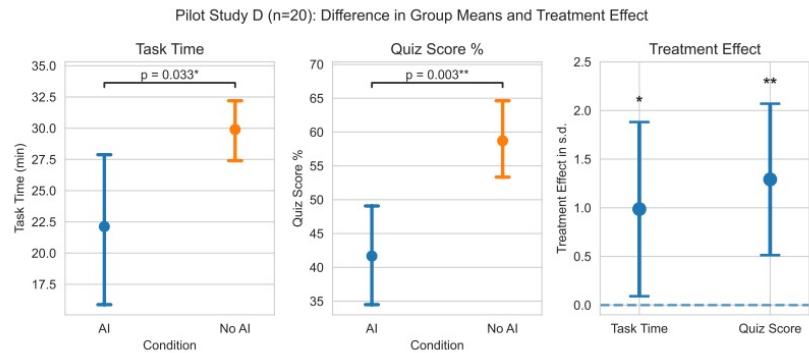


Figure 5: Difference in means of overall task time and quiz score between the control (No AI) and treatment (AI Assistant) groups in Pilot Study D. Error bars represent 95% CI. Significance values correspond to treatment effect. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

AI usage paradoxes:

- Most useful in places where we are already strong enough to spot the mistakes
- Best for tasks we could do ourselves but don't want to waste time on, but can harm our own learning when skipping necessary struggles

More on Hallucinations and Ground Truth

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.



Sam Altman

@sama

...

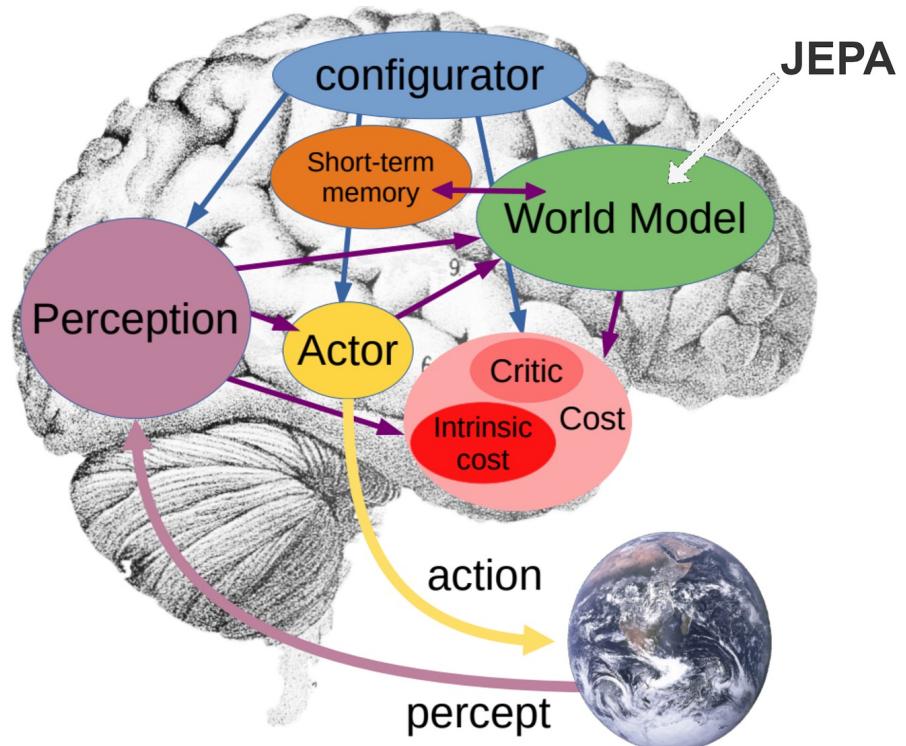
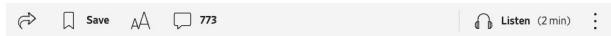
i am a stochastic parrot, and so r u

1:32 PM · Dec 4, 2022

World Model Approach

He's Been Right About AI for 40 Years. Now He Thinks Everyone Is Wrong.

Yann LeCun invented many fundamental components of modern AI. Now he's convinced most in his field have been led astray by the siren song of large language models.



The Platonic Representation Hypothesis

The Platonic Representation Hypothesis

Minyoung Huh^{*1} Brian Cheung^{*1} Tongzhou Wang^{*1} Phillip Isola^{*1}

Abstract

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the ways by which different neural networks represent data are becoming more aligned. Next, we demonstrate convergence across data modalities: as vision models and language models get larger, they measure distance between datapoints in a more and more alike way. We hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. We term such a representation the *platonic representation* and discuss several possible selective pressures toward it. Finally, we discuss the implications of these trends, their limitations, and counterexamples to our analysis.

Project Page: phillipi.github.io/prh
Code: github.com/minyoungg/platonic-rep

1. Introduction

AI systems are rapidly evolving into highly multifunctional entities. For example, whereas in the past we had specialized solutions for different language processing tasks (*e.g.*, sentiment analysis, parsing, dialogue), modern large

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

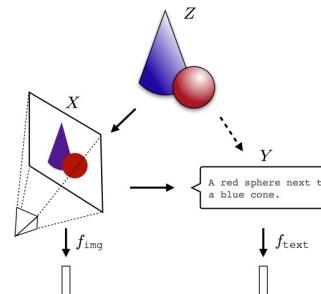


Figure 1. The Platonic Representation Hypothesis: Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z , and scaling model size, as well as data and task diversity, drives this convergence.

Evidence for the Platonic View?

Universally Converging Representations of Matter Across Scientific Foundation Models

Sathya Edamadaka^{1†}, Soojung Yang^{2*†}, Ju Li^{1,3},
Rafael Gómez-Bombarelli^{1*}

Abstract

Machine learning models of vastly different modalities and architectures are being trained to predict the behavior of molecules, materials, and proteins. However, it remains unclear whether they learn similar internal representations of matter. Understanding their latent structure is essential for building scientific foundation models that generalize reliably beyond their training domains. Although representational convergence has been observed in language and vision, its counterpart in the sciences has not been systematically explored. Here, we show that representations learned by nearly sixty scientific models, spanning string-, graph-, 3D atomistic, and protein-based modalities, are highly aligned across a wide range of chemical systems. Models trained on different datasets have highly similar representations of small molecules, and machine learning interatomic potentials converge in representation space as they improve in performance, suggesting that foundation models learn a common underlying representation of physical reality. We then show two distinct regimes of scientific models: on inputs similar to those seen during training, high-performing models align closely and weak

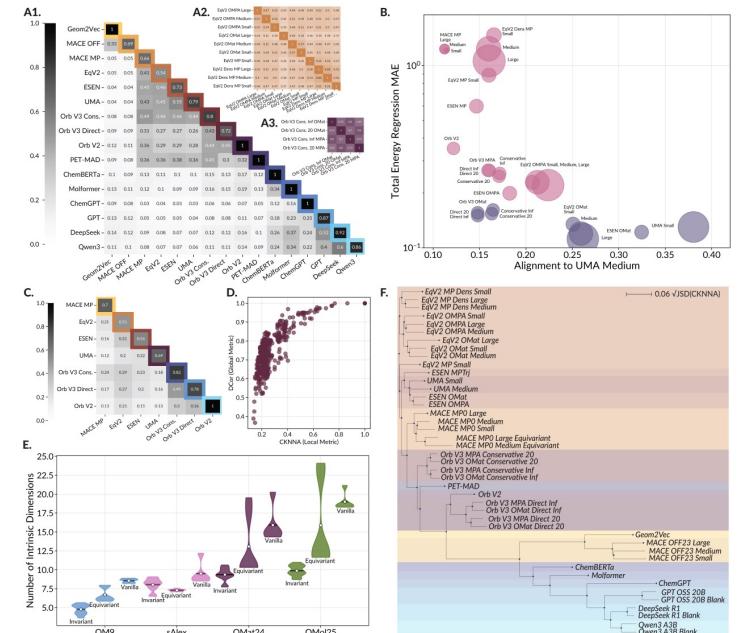


Fig. 1 Evidence for the convergence of scientific foundation model representations. A1 shows representational alignment between model architectures as measured by Centered Kernel Nearest-Neighbor Alignment (CKNNA) on structures from QM9 [17]. It is a condensed version of Fig. C6, as described in the Methods section, where each cell represents the average of several rows and columns. Some elements along the block diagonal are not 1 because we averaged several configurations

“Why do the language model and the vision model align? Because they’re both shadows of the same world”

Retrieval Augmented Generation (RAG)

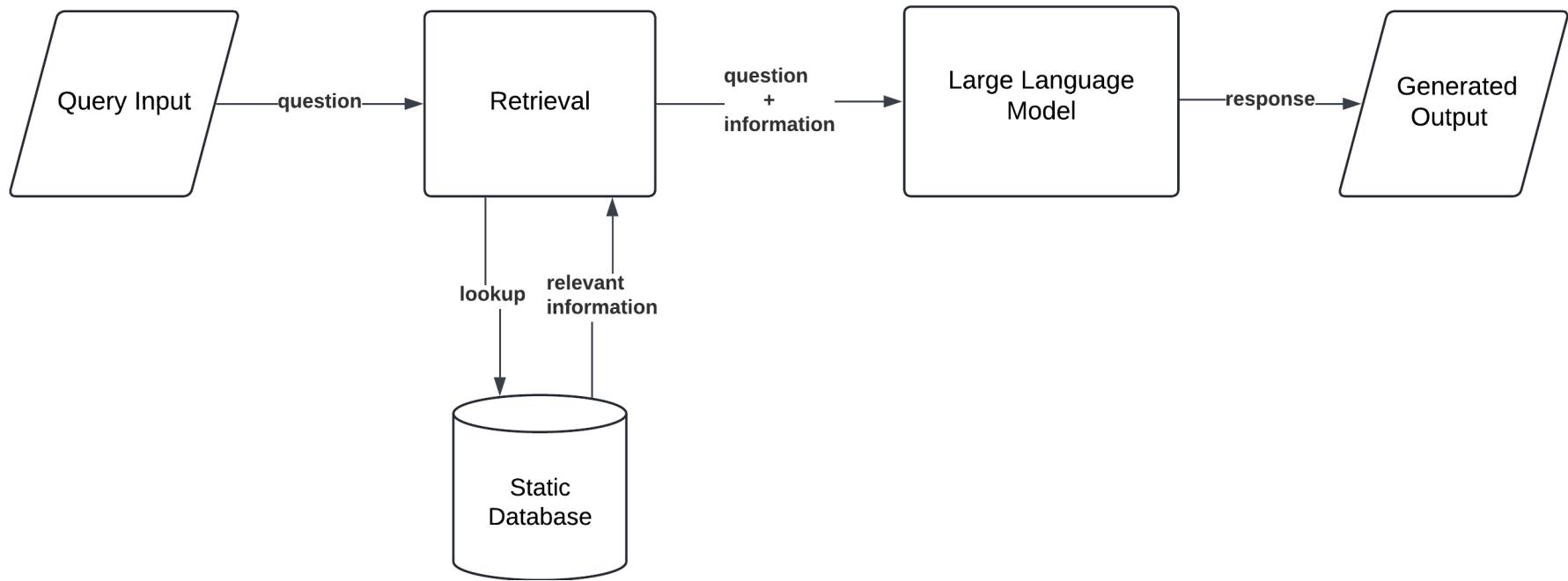
Retrieval-Augmented Generation (RAG)

- To solve the hallucination problem, and try to more closely map the model to reality, a commonly used approach is **Retrieval-Augmented Generation**, or RAG
- This is an architectural approach which restricts the search space to a pre-specified domain, helping to keep responses grounded in specific references
 - A key difference is that LLMs draw on their training material, while the reference corpus in RAG can be updated over time

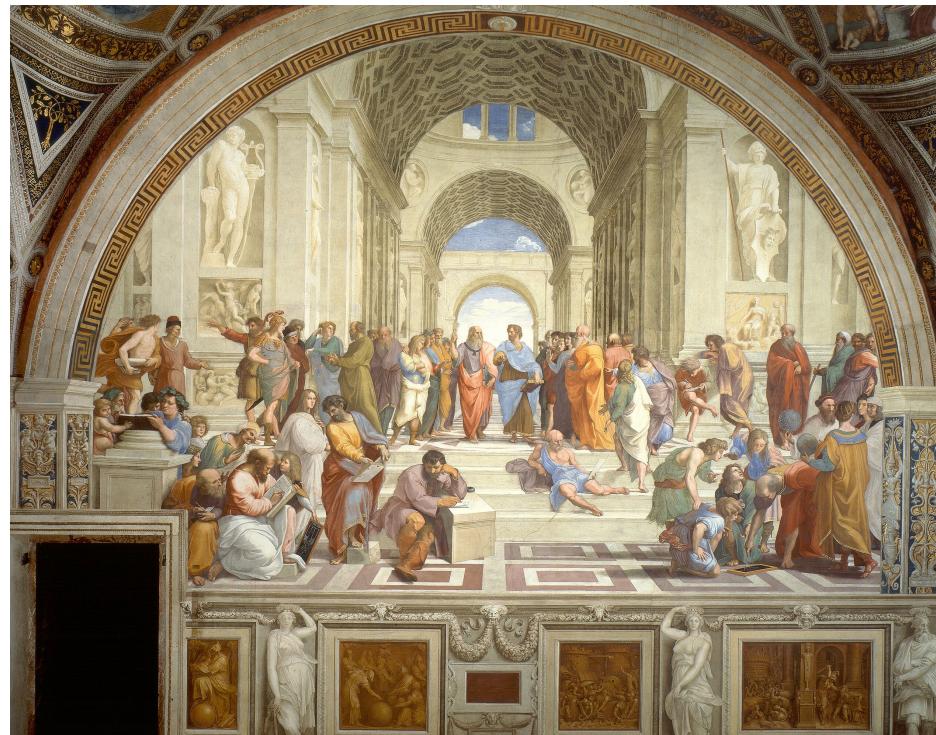
Basic RAG Approach

- Indexing
 1. Collect documents
 2. Process and clean content
 3. "Chunk" documents into pieces
 4. Index for future retrieval
- Retrieval
 1. User submits question
 2. Retrieve relevant chunks
 3. Pass chunk content to LLM
 4. Generate response

How RAG Works



Inferring Meaning from Relevant Context



RAG Design Tradeoffs

1. Document selection

- Sufficient to cover plausible range of questions; but too many reference documents dilutes retrieval and slows system

2. Chunking strategy

- Fixed-size, i.e. split at character or token limits
- Natural boundaries, i.e. paragraphs, sections, or document structure
- Semantic, keep ideas together through embedding similarity

3. Indexing Choice

- Keyword based (i.e., BM25; which uses a tf-idf [text frequency, inverse document frequency] based document ranking)
- Semantic (embeddings)

RAG Design Tradeoffs

4. Recall vs Precision

- More chunks = greater chance of “covering” the true answer
- ... At the cost of greater noise and cost

5. LLM Integration

- System prompt design
 - Instruct to cite sources, acknowledge uncertainty, prioritize retrieved information
- Context formatting
- How to manage empty retrieval

Case Study: Generative Regulatory Measurement

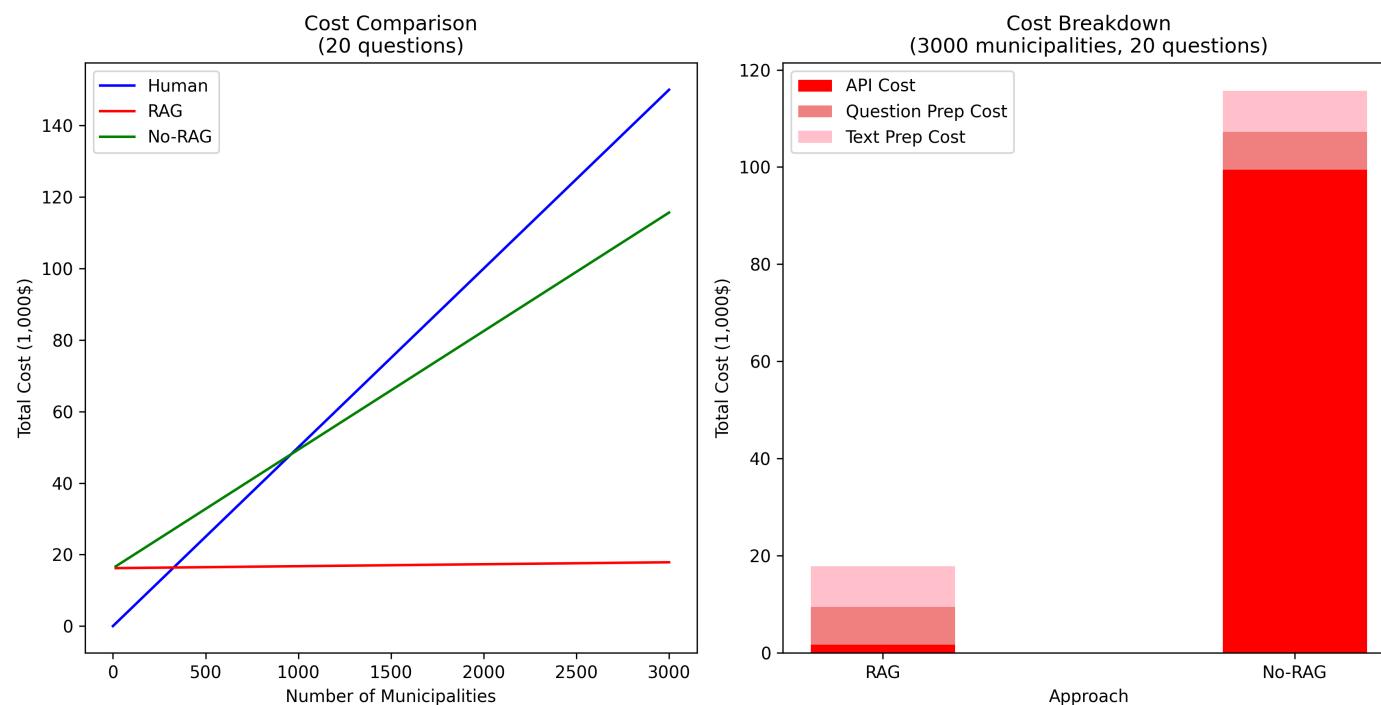
Zoning Regulations in U.S. Municipalities

- **Municipal Governance in the U.S.**
 - Municipalities are governed by state law
 - Includes various local government units, such as cities, towns, and villages
 - Zoning authority is a primary motivation for municipal incorporation
- **Scope of Zoning Regulations**
 - Land use regs.: define permitted uses for each zone
 - Bulk regs.: control bldg size, density, placement (e.g. height limits, setbacks, lot coverage)
 - Process rules: govern procedures to be followed when approving projects
 - Zoning codes are regularly updated to reflect changes in regulations and community needs
- **Collect data on broad range of US zoning regulations**
 - Data sources: American Legal Publishing, Municode, Ordinance.com
 - Coverage: Significant portion of US population

LLM Advantages and Disadvantages for Regulatory Analysis

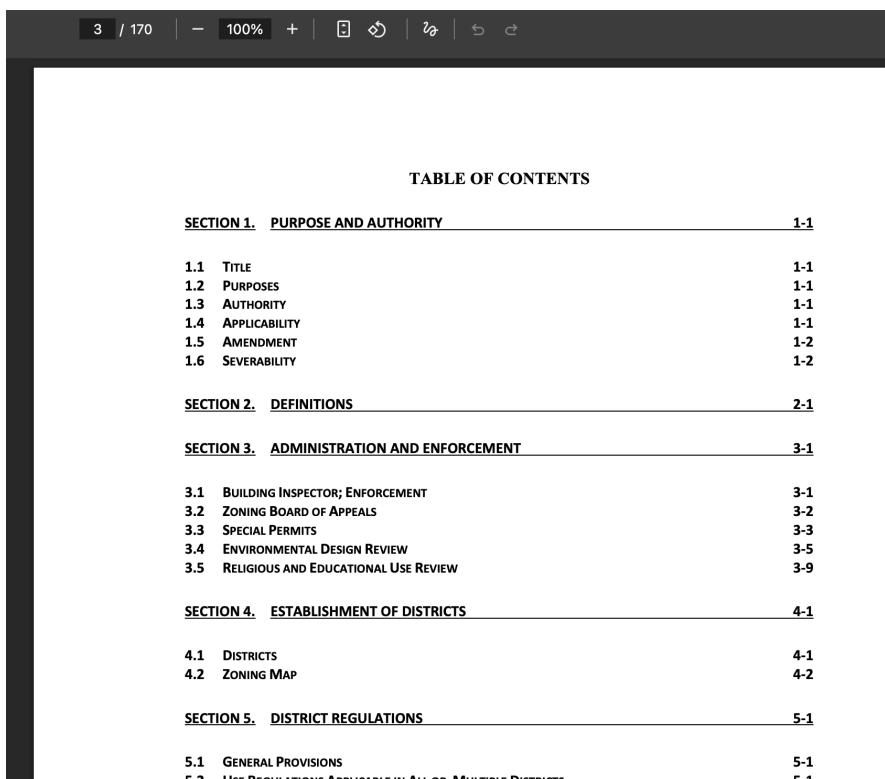
- Advantages
 - Substantially cheaper and faster than human analysis
 - Consistent interpretation of regulation—One LLM vs a team of human analysis
- Challenges
 - Hallucinations and inaccurate classifications
 - Solution: Train and validate against human classified data
 - Limited effective LLM context window
 - Use Retrieval Augmented Generation (RAG)

LLM vs Human Cost Analysis



Municipal Zoning Codes are Long

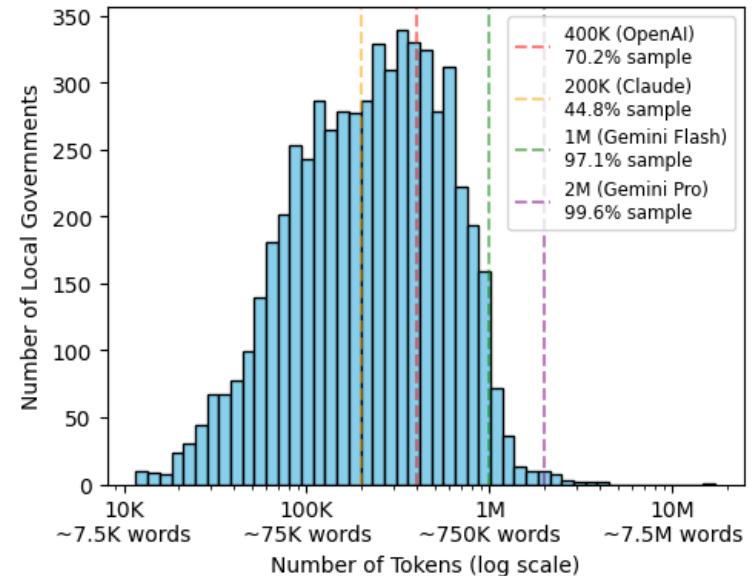
Arlington, MA Zoning Code



The screenshot shows the Table of Contents page of the Arlington, MA Zoning Code. The page includes a navigation bar at the top with icons for search, zoom, and orientation. The Table of Contents lists various sections and their page numbers:

TABLE OF CONTENTS	
<u>SECTION 1. PURPOSE AND AUTHORITY</u>	1-1
1.1 TITLE	1-1
1.2 PURPOSES	1-1
1.3 AUTHORITY	1-1
1.4 APPLICABILITY	1-1
1.5 AMENDMENT	1-2
1.6 SEVERABILITY	1-2
<u>SECTION 2. DEFINITIONS</u>	2-1
<u>SECTION 3. ADMINISTRATION AND ENFORCEMENT</u>	3-1
3.1 BUILDING INSPECTOR; ENFORCEMENT	3-1
3.2 ZONING BOARD OF APPEALS	3-2
3.3 SPECIAL PERMITS	3-3
3.4 ENVIRONMENTAL DESIGN REVIEW	3-5
3.5 RELIGIOUS AND EDUCATIONAL USE REVIEW	3-9
<u>SECTION 4. ESTABLISHMENT OF DISTRICTS</u>	4-1
4.1 DISTRICTS	4-1
4.2 ZONING MAP	4-2
<u>SECTION 5. DISTRICT REGULATIONS</u>	5-1
5.1 GENERAL PROVISIONS	5-1
5.2 USE REGULATIONS APPLICABLE IN ALL OR MULTIPLE DISTRICTS	5-1

Distribution of Municipal Code Lengths



Municipal Zoning Codes are Complex

Bulk Regulations Overview (Arlington, MA)

5-2 / DISTRICTS & USES

5.3 DIMENSIONAL REGULATIONS APPLICABLE IN ALL OR MULTIPLE DISTRICTS

5.3.1. Lot Area Per Dwelling Unit

Minimum lot area per dwelling unit shall control the maximum number of dwelling units, of all types, that can be constructed on contiguous land under one ownership in one zoning district. In the business (B) districts, where a lot may contain both residential and nonresidential principal structures, the maximum number of dwelling units is computed by dividing the total land area by the minimum lot area per dwelling unit. Land in lower density districts used for buildings in higher density districts shall not be included in the calculation of minimum lot area per dwelling unit for dwellings in the higher density district.

5.3.2. Reduction of Lot Areas and Separation of Lots

- A. The lot or yard areas required for any new building or use may not include any part of a lot that is required by any other building or use to comply with any provisions of this Bylaw, nor may these areas include any property of which the ownership has been transferred after the effective date of this Bylaw if the property was a part of the area required for compliance with the dimensional regulations applicable to the lot from which such transfer was made.
- B. Lots separated or transferred in ownership must upon transfer either comply with the provisions of this Bylaw or be deemed noncompliant.

5.3.3. Spacing of Residential and Other Buildings on One Lot

- A. Where two or more main buildings to be used as dwellings are proposed for construction on property in one ownership or where one or more of the buildings are proposed on land where there are one or more existing residential buildings, the required front, side, and rear yards shall be provided between each building and assumed lot lines shown upon the building permit application. However, the Board of Appeals or, for projects requiring Environmental Design Review, the Arlington

Bulk Regulations Table (Arlington, MA)

District Regulations / 5-17

R District Building Height and Floor Area Ratio Regulations (see 5.4.2(B) for exceptions)

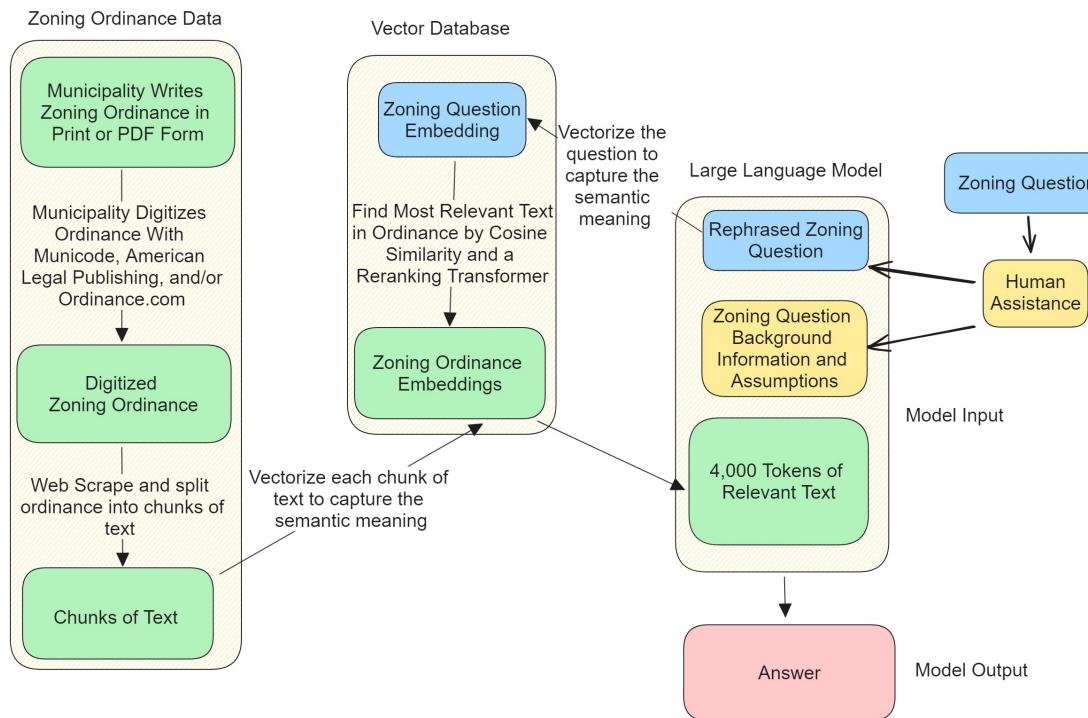
District Use	Maximum Allowed		
	Maximum Height (ft.)	Maximum height (stories)	Maximum Floor Area Ratio (FAR)
R0, R1			
Single Family detached dwelling	35	2 ½
Other permitted structure	35	2 ½	0.35
R2			
Single family detached dwelling, two-family dwelling or duplex dwelling	35	2 ½
Other permitted structure	35	2 ½	0.35
R0, R1, R2			
Accessory structures (> 80 sq. ft.) and private garages	20	2
Minor accessory building (<= 80 sq. ft.)	7	1
R3			
Principal building or structure	35	3	0.75
Detached accessory structure (> 80 sq. ft.)	20	2
Detached accessory structure (<= 80 sq. ft.)	7	1
R4			
Single-family detached dwelling, two-family dwelling, duplex dwelling	35	2 ½
Three-family dwelling	35	3
Townhouse structure	35	3	0.70
Apartment conversion	40	3
Nursing home, dormitory, or single-room occupancy conversion	35	3	0.70-
Any other permitted structure	35	2 ½	0.35
Detached accessory structure (> 80 sq. ft.)	20	2
Detached accessory structure (<= 80 sq. ft.)	7	1
R5			
Any residential or other principal structure	35	2	0.80

DISTRICTS & USES

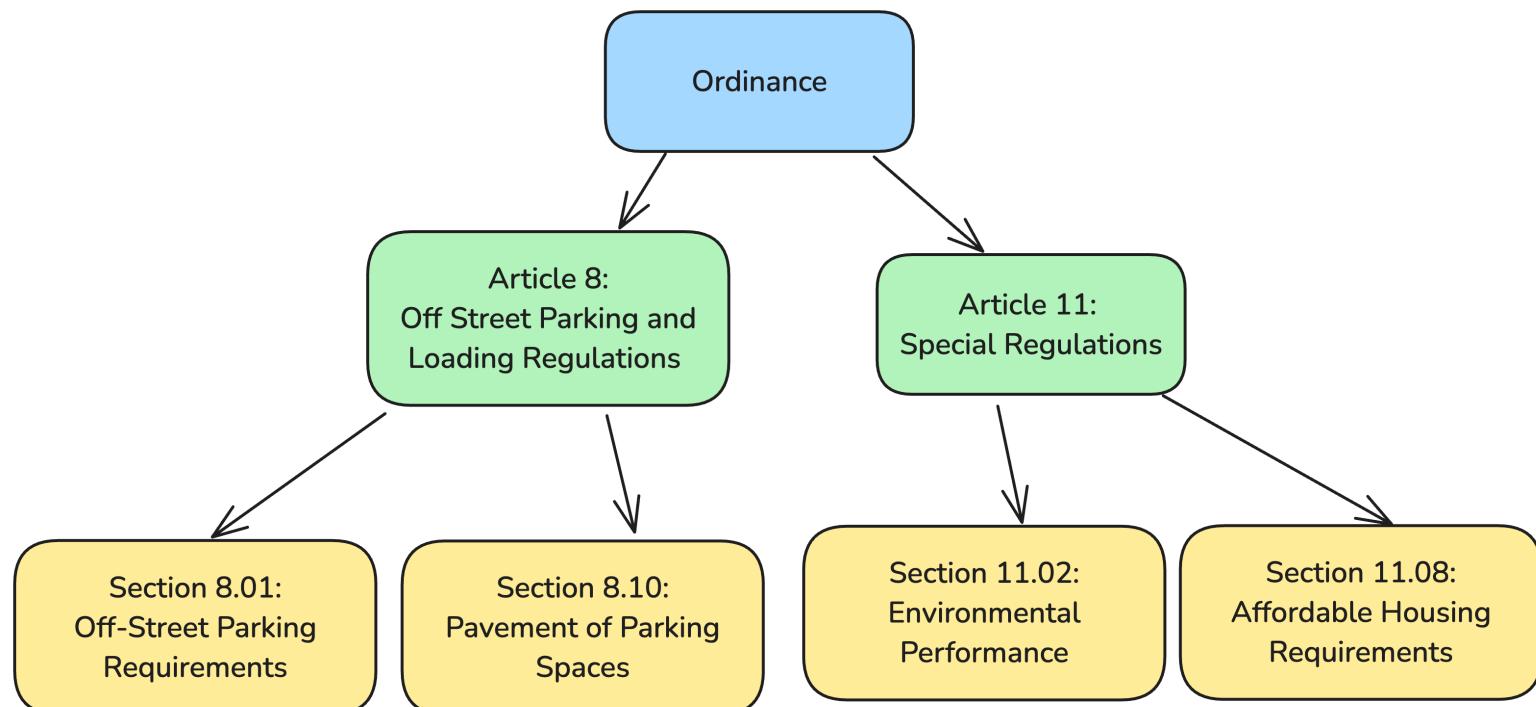
Estimation, Training, and Testing

- 14 questions from Pioneer Institute (Glaeser Ward 09): + 6 questions for minimum lot sizes, affordable housing mandates, & housing process questions
 - Developed process on 107 municipalities
 - Tested on 30 municipalities (only once)
 - Reserved 50 municipalities for potential fine-tuning (not included in the current study)
- Classification process involves:
 - Multi-step processing
 - Subtasks
 - Double-checking
 - Prompt engineering
 - Adding question specific background and assumptions

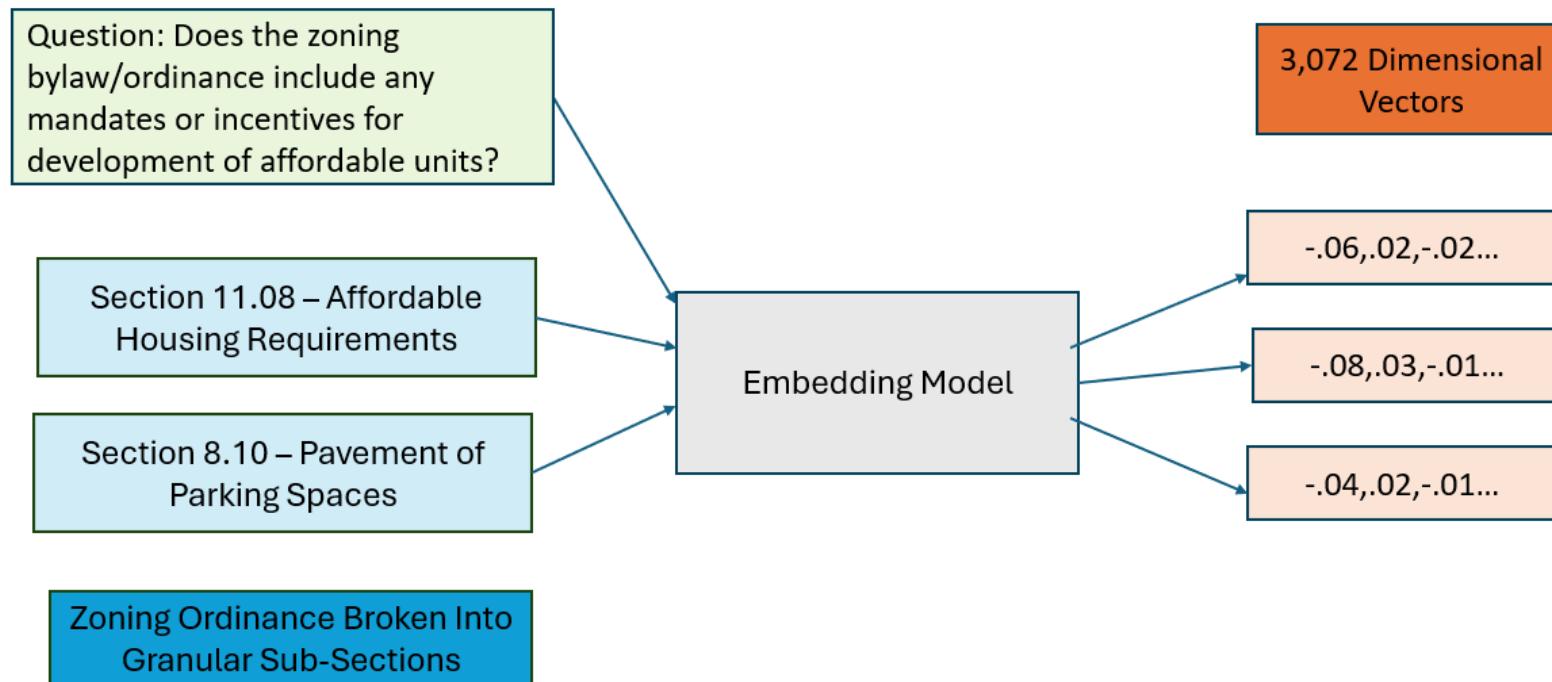
Processing Municipal Codes Using RAG + LLMs



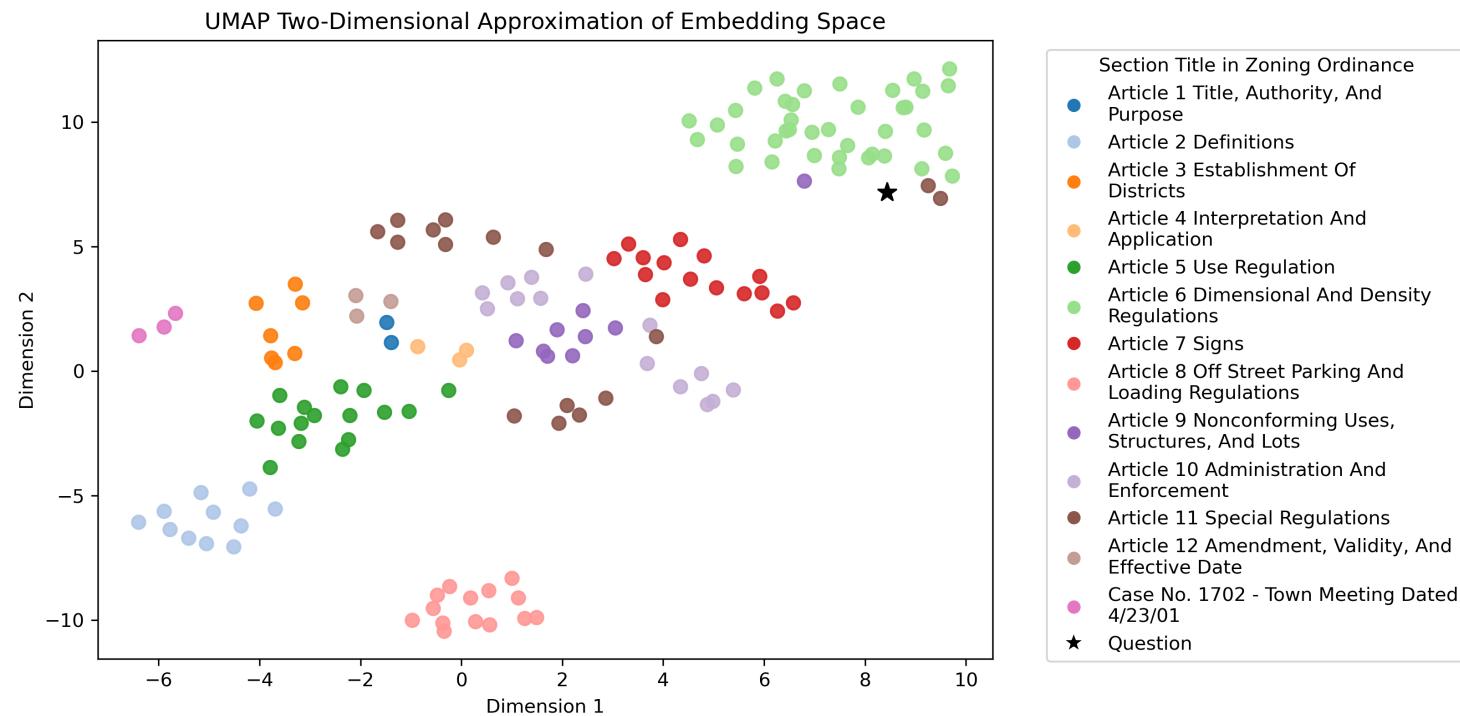
Example: First Split Up Ordinance Based on Hierarchical Structure (Arlington, MA)



Next, We Embed Text and a Question for Arlington, MA

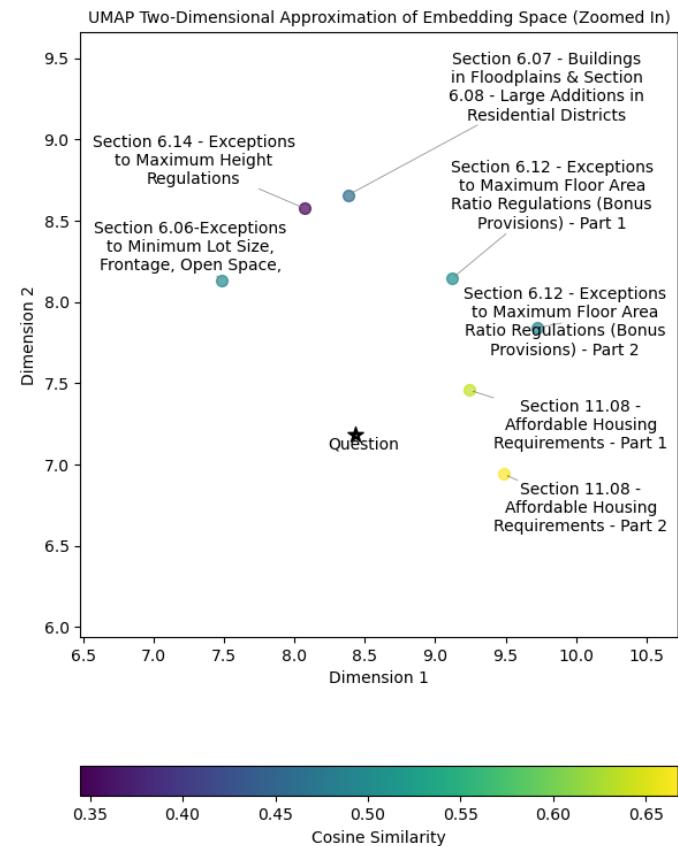


Embedding Space Clusters Similar Text



Relevant Text Close in Space to Question

- We measure the distance from the question vector to the text vector using normalized cosine-similarity (in n-dimensional space)
- Relevant sections on affordable housing and exceptions to lot size and floor area requirements are near the question vector
- We also use a re-ranking algorithm on high cosine-similarity text



We Provide the LLM With Question Background Information

- **Background Information:** Inclusionary zoning requires or encourages developers to include affordable dwelling units within new developments of market rate homes. Some municipalities call it “incentive zoning” ... (omitted for brevity)
- **Assumptions:** Do not include provisions for entirely affordable, subsidized housing development by public or non-profit corporations. Also do not include provisions under “rate of development” headings that exempt affordable units from project phasing and growth caps.

Relevant Text Enables A Correct Answer

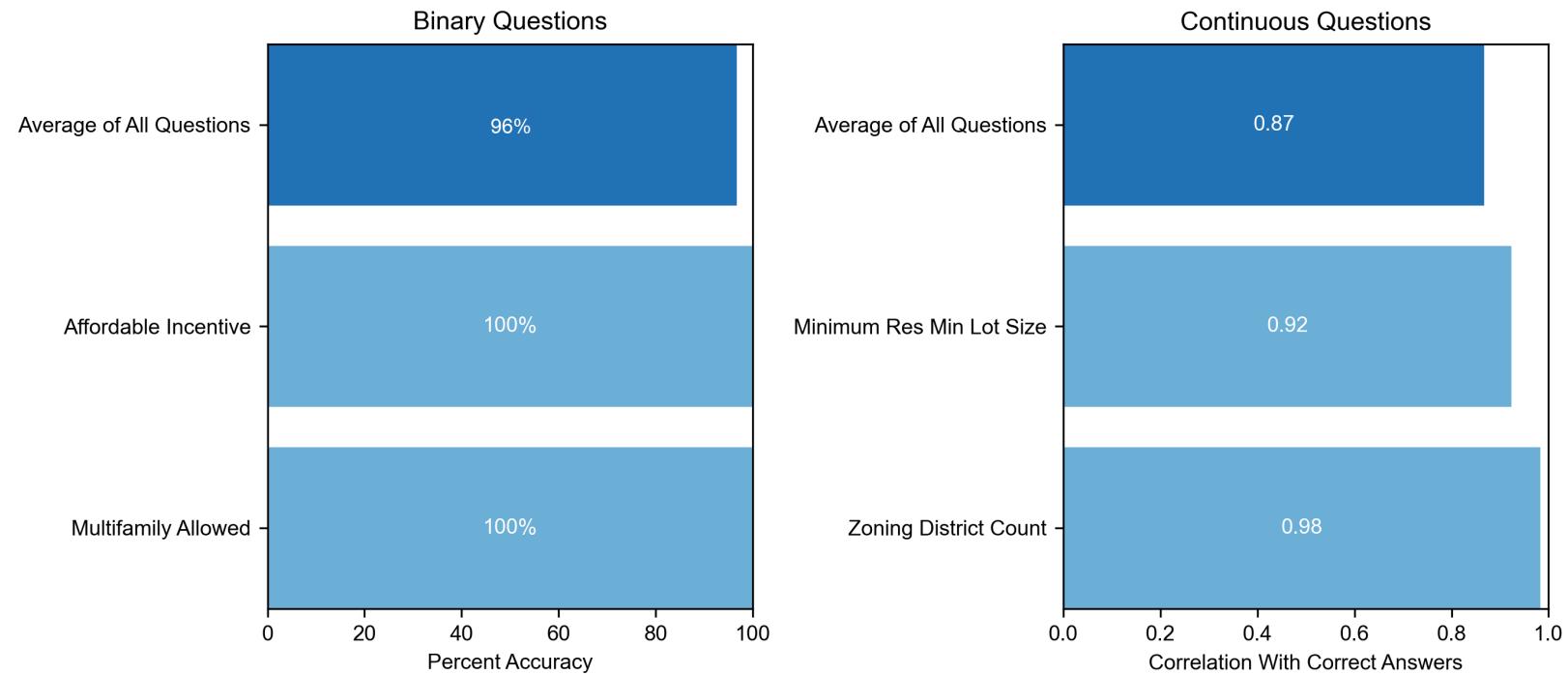
- **GPT-4 Turbo Output:** The zoning bylaw/ordinance provided does indeed include mandates and incentives for the development of affordable units. Here's a breakdown of the evidence supporting this conclusion:
 1. ****Mandates for Affordable Housing Development**:** The ordinance mandates that 15% of the residential units in new projects be affordable units (Section 11.08, Part 1, d. REQUIREMENTS, 1.). This requirement applies to all new residential projects, including phased or segmented developments, with six or more units subject to Environmental Design Review (Section 11.08, Part 1, b. APPLICATION). The affordable units must conform to specific affordability criteria, ensuring they are accessible to households earning no more than 60% of the median income for rental units and no more than 70% of the median income for homeownership units (Section 11.08, Part 1, c. DEFINITIONS).

(other evidence omitted for brevity)

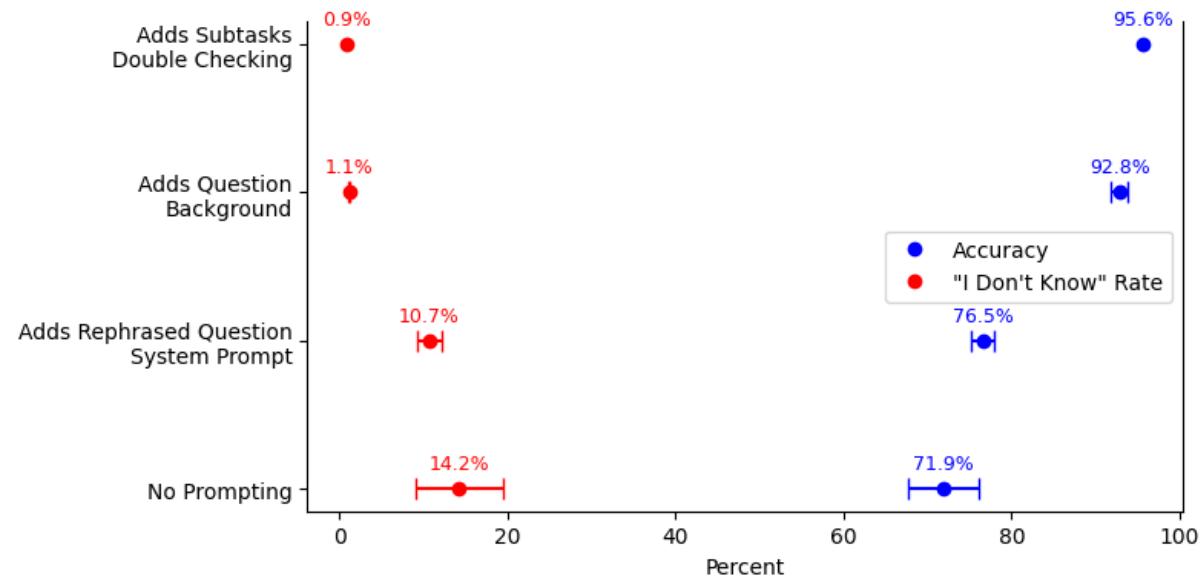
Given the evidence from the zoning bylaw/ordinance, it is clear that the town has both mandates and incentives in place for the development of affordable housing units.

ANSWER: YES

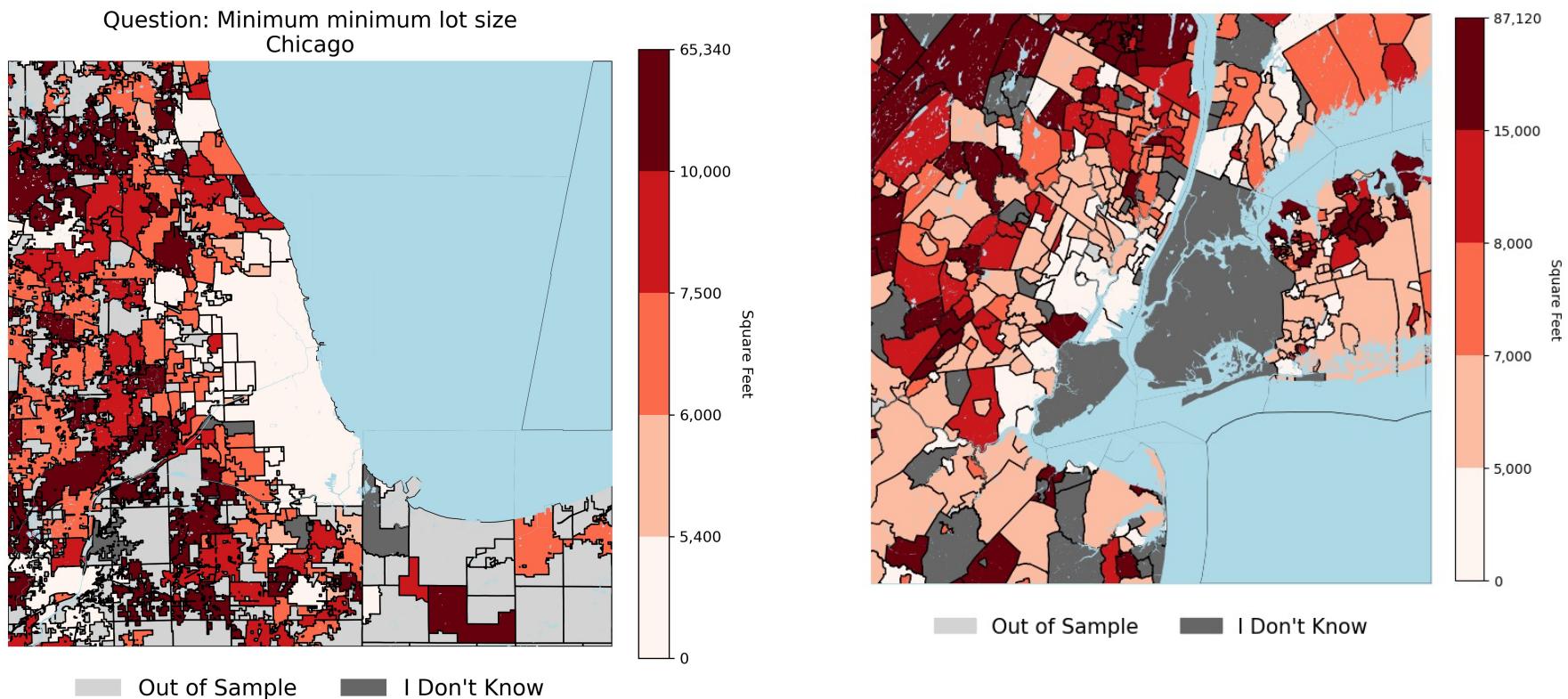
High Accuracy on Binary Questions and Continuous Measures



Adding Question Background Info Most Important Training Step



Example of Output: Min Lot Sizes



AI Applications

RAG Examples



- AI co-pilot for doctors based on validated clinical findings



- AI for Law



- AI as an answer engine

Where LLMs and RAG Add Value in Finance

1. **Text-Heavy Workflows:** Research synthesis, compliance review, client communications, document Q&A
2. **Analyst Co-Pilot:** Report drafting, coding assistance, document navigation, data exploration
3. **Frontier Challenges:** Fully automated decisions, high-stakes autonomy, non-stationary data, real-time trading