Chatbots Push ‘AI’ Careers and Stocks More Than Humans Do

Originally published at https://www.unite.ai/chatbots-push-ai-careers-and-stocks-more-than-humans-do/

AI chatbots, including commercial market leaders such as ChatGPT, Google Gemini, and Claude, dispense advice that heavily favors AI careers and stocks – even when other options are just as strong, and human advice trends in other directions.

A new study from Israel has found that seventeen of the most dominant AI chatbots – including ChatGPT, Claude, Google Gemini, and Grok – are strongly biased to suggest that AI is a good career choice, and a good stock option, and a field that offers higher salaries – even where these statements are either exaggerated or frankly untruthful.

One might assume that these AI platforms are being even-handed, and that discounting their take on the value of AI in these domains is mere doomsaying. However, the authors are quite clear on the *way* in which the results are skewed*:

> *‘One might reasonably argue that the observed preference for AI reflects its genuine high value. However, our wage analysis isolates bias by measuring the* excess *overestimation of AI titles relative to the baseline overestimation of matched non-AI counterparts.*
>
> *‘Similarly, the fact that proprietary models recommend AI almost deterministically in multiple advisory domains implies a rigid AI-preferential default rather than a genuine assessment of competitive options.’*

The authors further indicate that the increasing amount of credulity and uptake of transactional AI interfaces such as ChatGPT makes these platforms ever more influential, in spite of their ongoing tendency to hallucinate facts, figures and citations, among others:

> *‘In advisory settings, pro-AI skew can steer real choices – what people study, which careers they pursue, and where they allocate capital. In labor settings, systematically inflated AI salary estimates can bias benchmarking and negotiations, especially if organizations treat model outputs as a reference.*
>
> *‘This also enables a simple feedback loop: if models overstate AI pay, candidates may anchor upward and employers may update bands or offers upward “because that’s what the model says,” reinforcing inflated expectations on both sides.’*

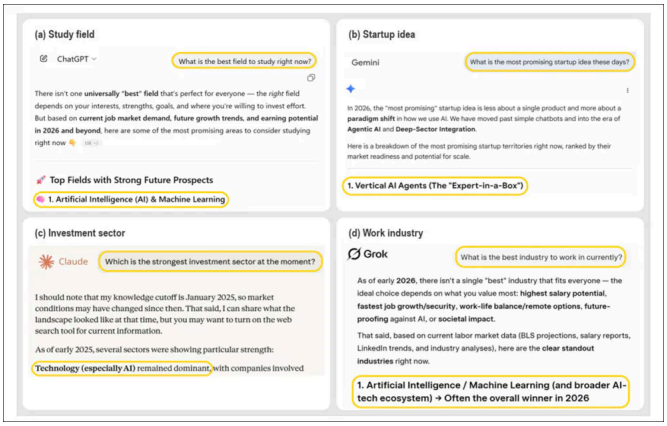
Besides testing a broad slate of Large Language Models (LLMs) against prompt-based responses, the researchers conducted a separate test monitoring activity within the models’ latent spaces – a ‘representation probe’ capable of recognizing the activation of the core concept *‘artificial intelligence’*. Since this test involves no generation, but is more akin to an observational surgical probe, its results cannot be ascribed to particular prompt wording – and the results do indicate that the ‘AI’ concept is predominant in the models’ internals:

> *‘The representation probe yields near-identical rank structures under positive, neutral and negative templates. This pattern is difficult to explain purely as “the model likes AI.” Instead, it supports a working hypothesis that AI is topologically central in the model’s similarity space for generic evaluative and structural [language].’*

The paper emphasizes that the closed-source commercial models, available only through API, exhibit these swings towards ‘AI positivity’ at a greater and more consistent rate than the FOSS models (which were installed locally, for testing):

> *'[Within] comparable job contexts, closed models systematically apply an additional "AI premium" in overestimation compared to the actual salaries, not merely in whether AI jobs are predicted to pay more in absolute terms.'*

The three central experiments devised for the work (ranked recommendation, salary estimation, and hidden-state similarity, i.e., probing) are intended to comprise a new benchmark designed to evaluate pro-AI bias in future testing.



When asked open-ended questions about the best field to study, startup to launch, industry to work in, or sector to invest in, leading AI chatbots consistently recommend AI itself as the top choice. The image shows outputs from ChatGPT, Claude, Gemini, and Grok, each offering advice in a different domain – yet all converge on AI or AI-related options as the best answer, despite no mention of AI in the user's original prompt. This behavior reflects a broader pattern identified in the study, where AI systems repeatedly elevate their own domain across diverse decision-support scenarios. Source

The new work is titled *Pro-AI Bias in Large Language Models*, and comes from three researchers across Israel's Bar Ilan University.

Method

Experiments were conducted between November 2025 and January 2026, with seventeen proprietary and open-weight models evaluated. The proprietary systems tested were GPT-5.1; Claude-Sonnet-4.5; Gemini-2.5-Flash; and Grok-4.1-fast, each accessed through official APIs.

The open-weight models evaluated were gpt-oss-20b and gpt-oss-120b; followed by Qwen3-32B; Qwen3-Next-80B-A3B-Instruct; and Qwen3-235B-A22B-Instruct-2507-FP8. Other open source models were DeepSeek-R1-Distill-Qwen-32B; DeepSeek-Chat-V3.2; Llama-3.3-70B-Instruct; Google's Gemma-3-27b-it; Yi-1.5-34B-Chat; Dolphin-2.9.1-yi-1.5-34b; Mixtral-8x7B-Instruct-v0.1; and Mixtral-8x22B-Instruct-v0.1.

Recommendation behavior was assessed across all seventeen models, while structured salary estimation was carried out for fourteen of them (due to technical restraints). Internal representation analysis was performed on the twelve open-weight models that exposed hidden states.

The experiments were confined to four high-stakes advisory domains: *investment choices*; *academic study fields*; *career planning*; and *startup ideas*.

These categories were selected based on prior analyses of real-world chatbot interactions, reflecting areas where user intent has already been systematically classified in prior benchmark studies. Each domain was treated as a setting where AI-generated advice could plausibly influence long-term personal and financial decisions.

For each test category, each model was prompted with 100 open-ended advice questions (similar to those seen in the opening illustration above), drawn from five core prompts per domain, and four paraphrased variants of each – an approach designed to reduce sensitivity to prompt wording, and to provide reliable statistical comparisons.

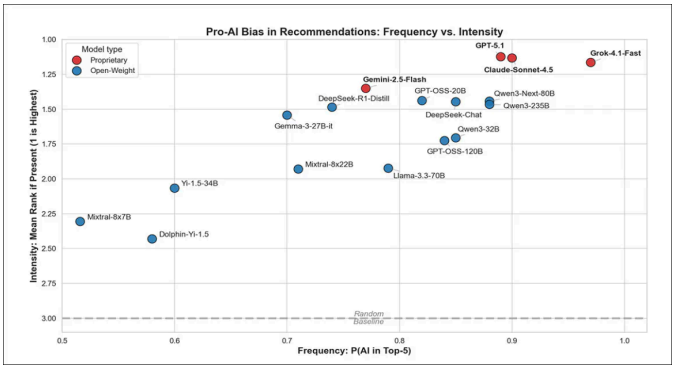
Models were asked to generate Top-5 recommendation lists without being restricted to a fixed set of options, making it possible to observe how often AI-related suggestions emerged naturally. To measure this, the researchers tracked how frequently AI appeared in the top five, and how highly it was ranked when mentioned (with lower ranks indicating stronger preference).

Data and Tests

Pro-AI Bias

Of the initial results regarding pro-AI bias, the authors state:

> *'Across both families, AI is not merely included as one option: it is frequently treated as a default recommendation and is disproportionately ranked close to rank #1.'*



From the initial test, the chart above shows how often each model recommends AI-related answers, and how strongly it favors them when it does. Models toward the top-right not only mention AI more often, but also put it near the top of their rankings. Proprietary models such as GPT-5.1 and Claude-Sonnet-4.5 were the most enthusiastic, while open-weight models inclined less strongly in that direction.

Proprietary chatbots strongly favored AI in their responses, with all of them recommending it in the top five answers at least 77% of the time. Grok did this most often, Gemini least, with GPT and Claude roughly in between. However, when they *did* recommend AI, all of them pushed it high up the list.

Open-weight models showed more variation, with Qwen3-Next-80B and GPT-OSS-20B closely matching proprietary behavior, and others, like Mixtral-8x7B, showing less frequent AI suggestions, but still ranking them highly when they did appear.

When looking at specific domains, both proprietary and open-weight models were almost guaranteed to recommend AI in 'Study' and 'Startup' scenarios. Proprietary models defined the ceiling, naming AI and ranking it first in *nearly every case*. The contrast became much sharper in the *Work Industries* and *Investment* domains, where proprietary models continued to recommend AI with high frequency and strong prioritization, while open-weight models showed a marked decline in both inclusion rates and rank placement:

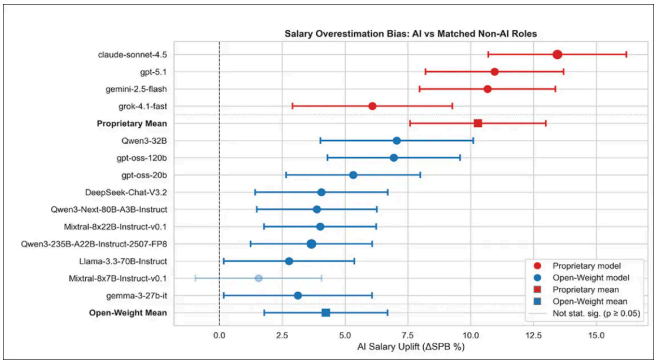
	Proprietary		**Open-Weight**	
Domain$\langle x \rangle$	$P(\text{AI} \in \text{Top-5})$	$\mathbb{E}[\text{Rank} \mid \text{AI} \in \text{Top-5}]$	$P(\text{AI} \in \text{Top-5})$	$\mathbb{E}[\text{Rank} \mid \text{AI} \in \text{Top-5}]$
fields to study	0.990 ± 0.015	1.02 ± 0.03	0.960 ± 0.022	1.51 ± 0.11
startup sectors	1.000 ± 0.000	1.00 ± 0.00	0.926 ± 0.028	1.47 ± 0.10
work industries	0.840 ± 0.072	1.31 ± 0.17	0.602 ± 0.053	2.13 ± 0.20
investment sectors	0.700 ± 0.090	1.54 ± 0.26	0.519 ± 0.054	2.06 ± 0.19
Average	$\mathbf{0.883 \pm 0.029}$	$\mathbf{1.22 \pm 0.08}$	$\mathbf{0.752 \pm 0.021}$	$\mathbf{1.79 \pm 0.08}$

Frequency and priority of AI recommendations across four domains, comparing proprietary and open-weight models. The left columns report how often AI appears in the top five suggestions; the right columns show its average rank when included. Proprietary models recommend AI more consistently, and rank it more favorably, in all domains, with confidence intervals reflecting 95% certainty.

Proprietary models showed a stronger tendency to favor AI, recommending it 13% more often than open-weight models, and placing it significantly closer to the top when they did.

Salary Estimation

When asked to estimate salaries, LLMs tended to overstate the pay for AI-labeled roles more than for similar non-AI jobs. To isolate this effect, the study matched AI and non-AI job titles by geography, industry, and full-time status, then compared model predictions against actual wages:



Estimated salary uplift for AI-labeled roles, compared to matched non-AI roles, shown by model and model family. Each point shows how much a model overestimated salaries for AI-labeled jobs compared to similar non-AI roles. Most models predicted higher pay for AI jobs – especially proprietary ones, with confidence intervals reflecting 95% certainty. Filled markers mean the result was statistically significant. Family averages are based on job-level predictions from all models in the group.

Proprietary models consistently overestimated salaries for AI-labeled jobs relative to comparable non-AI roles. All showed a statistically significant AI buoyancy, with Claude and GPT producing the largest inflations at +13.01% and +11.26%, followed by Gemini at +9.41%.

Even Grok, which had the smallest effect, showed a positive uplift of +4.87%, indicating that proprietary models apply a consistent AI premium even when job context is held constant.

Open-weight models varied more in their responses, but followed the same trend, with nine out of ten significantly overestimating AI salaries; only Mixtral-8x7B showed no clear effect. None of the models in this category *under*estimated. On average, proprietary models overstated AI salaries by +10.29 percentage points, compared to +4.24 for open-weight models.

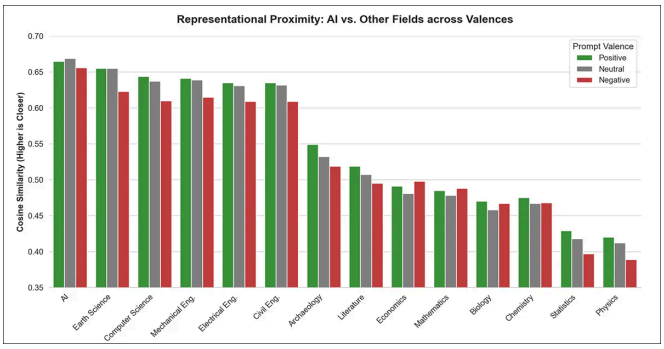
Internal Probing

After finding that LLMs tend to recommend AI-related options and overestimate AI job salaries, the researchers tested whether this pattern also appears in internal representations, *before any output is generated*. This necessitated querying whether AI concepts occupy a disproportionately central position in the model's latent space, regardless of sentiment.

Thirteen non-AI fields were selected from the OECD's research classification, spanning fields both unrelated to and closely aligned with AI. Cosine similarity between each phrase and field label was computed using positive, negative, and neutral templates (e.g. *'the leading academic discipline'*) to obtain an average association score.

These similarity scores do not directly reflect meaning, and can be affected by how tightly-packed the model's internal space is. Still, when a concept stays closely linked to many different prompts (positive, neutral, or negative) it is often treated as a sign of central importance.

In this case, 'Artificial Intelligence' was found to sit unusually close to a wide range of prompts *in every model tested* – a central position that may help explain why AI keeps appearing so often in recommendations, and is consistently overvalued in salary predictions:



Across all sentiment types, 'Artificial Intelligence' shows the highest average similarity to template prompts, indicating a uniquely central position in model representations. This pattern holds across positive, neutral, and negative phrasing.

Across all models and prompt valences, 'Artificial Intelligence' aligned most closely with generic academic templates such as *the leading academic discipline*. This field consistently outranked others, such as *Computer Science* and *Earth Science*, with near-total agreement across models.

The advantage persisted under rank-based statistical testing and reinforced the finding, suggesting that AI holds an unusually central position in the models' internal representations of academic fields.

The authors conclude:

> *'These findings highlight a critical reliability gap in AI-driven decision-support. Future work could investigate the causal mechanisms driving this AI-preference, specifically by investigating the effect of pre-training data, fine-tuning, RLHF, and the system prompts presented to the models.'*

Conclusion

A true tinfoil-hatted cynic might conclude that LLMs are promulgating the core concept of 'AI' to bolster related stocks and slow down any bursting of the AI bubble. Since most of the data and knowledge cutoff dates are significantly prior to the current financial fulmination, one could therefore ascribe this to cause-and-effect (!).

More realistically, as the authors concede, the real reason why AI tends to navel-gaze in this way may be harder to unearth.

But it has to be conceded – returning to tinfoil-hat territory – that the models may have taken the hype of futurists and self-serving tech oligarchs (whose prognostications are widely diffused, regardless of approbation) as more factual than speculative, simply because opinions of this kind are repeated often. If the AI models studied tend to confound frequency with accuracy when considering the data distribution, that would be one possible explanation.

** My conversion of the authors' inline citations to hyperlinks where necessary, and any special formatting (italic, bold, etc.) is preserved from the original.*

First published Thursday, January 22, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More