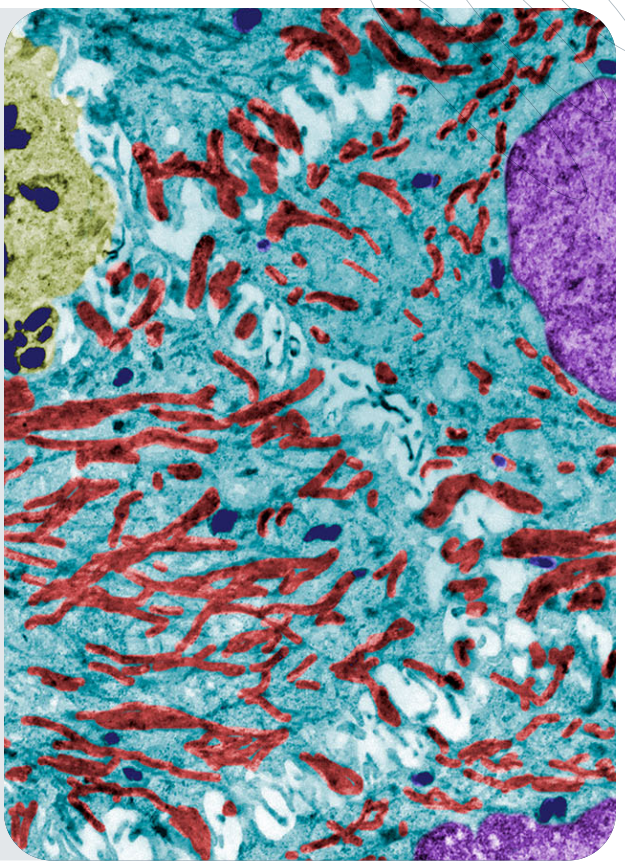
SCITARIS
BIOPHARMA STRATEGY



Scitaris Original Research

Can LLMs judge in-licensing opportunities? **Benchmarking prompt-engineered foundation models for biopharma strategy**

Study Objective

This study benchmarks three large language models (LLMs) — Copilot, Claude, and Gemini — evaluating their ability to provide accurate and actionable recommendations for critical decision-making in drug development. Additionally, we assess whether prompt engineering can enhance baseline performance.

Methodology

We used the Researcher agent in Copilot, Claude Sonnet 4.5 with extended thinking enabled, and Gemini 3 in this benchmark study. We validated model outputs against an internal ground truth dataset comprising 40 independent analyses of drug in-licensing opportunities conducted by a team of PhD-level biopharma strategists. Each analysis applied a systematic framework to score opportunities across eleven critical dimensions, e.g., pharmacokinetics (PK), pharmacodynamics (PD), efficacy, safety, and differentiation, condensed into a single recommendation score ranging from 1 (*not recommended*) to 5 (*highly recommended*). Initial tests prompted both models to rate in-licensing opportunities of drugs for specific indications on a 1–5 scale. Responses were manually collected and compared to the ground truth.

We further evaluated prompt engineering by augmenting the base prompt with:

- Sub-prompts with explicit instructions for each of the eleven subdimensions, incl. assignment of an expert role for a more defined context
- Clear specification of dimension-specific rating criteria (1–5 scale)
- A “referee” step consolidating dimension scores into a final recommendation with rationale

Results

Baseline Performance (Figure 1)
Neither Copilot, Claude, nor Gemini demonstrated meaningful correlation with ground truth.

- **Copilot overall misjudged 77% of opportunities.** It exhibited a strong bias toward positive ratings, with 39% of weak opportunities incorrectly suggested for in-licensing. Its lowest score was 3 out of 5 (observed in 4 assets).
- **Claude misjudged 64% of opportunities.** Its recommendation scores were more evenly spread across the 1–5 scale than Copilot's or Gemini's. Still, not a single extreme score of 1 or 5 was assigned. Further, Claude's ratings correlated neither with ground truth nor with any other tested model, resulting in numerous false positive and false negative recommendations.
- **Gemini misjudged 67% of opportunities.** Like Copilot, it typically rated opportunities more favorably than ground truth. However, its scores showed greater variability across opportunities, with a lowest score of 2 out of 5 (observed in 3 assets).

Next, we tested the engineered prompt. While formatting and presentation of output improved significantly, predictive accuracy did not. Overall, models tend to avoid extreme ratings reducing usability in decision making process. Furthermore, the few critical scores given by the models did not correlate with each other.

However, it wasn't that models were unable to provide any valid insights. In obvious cases, such as rating the in-license opportunity of a drug in a clearly unfavorable indication (e.g., oncology ADCs in non-oncology HPV infection or neuropathic pain), the model's recommendations were correct.

This suggests that models currently lack the ability to incorporate nuanced details and evaluate them properly for critical decision-making.

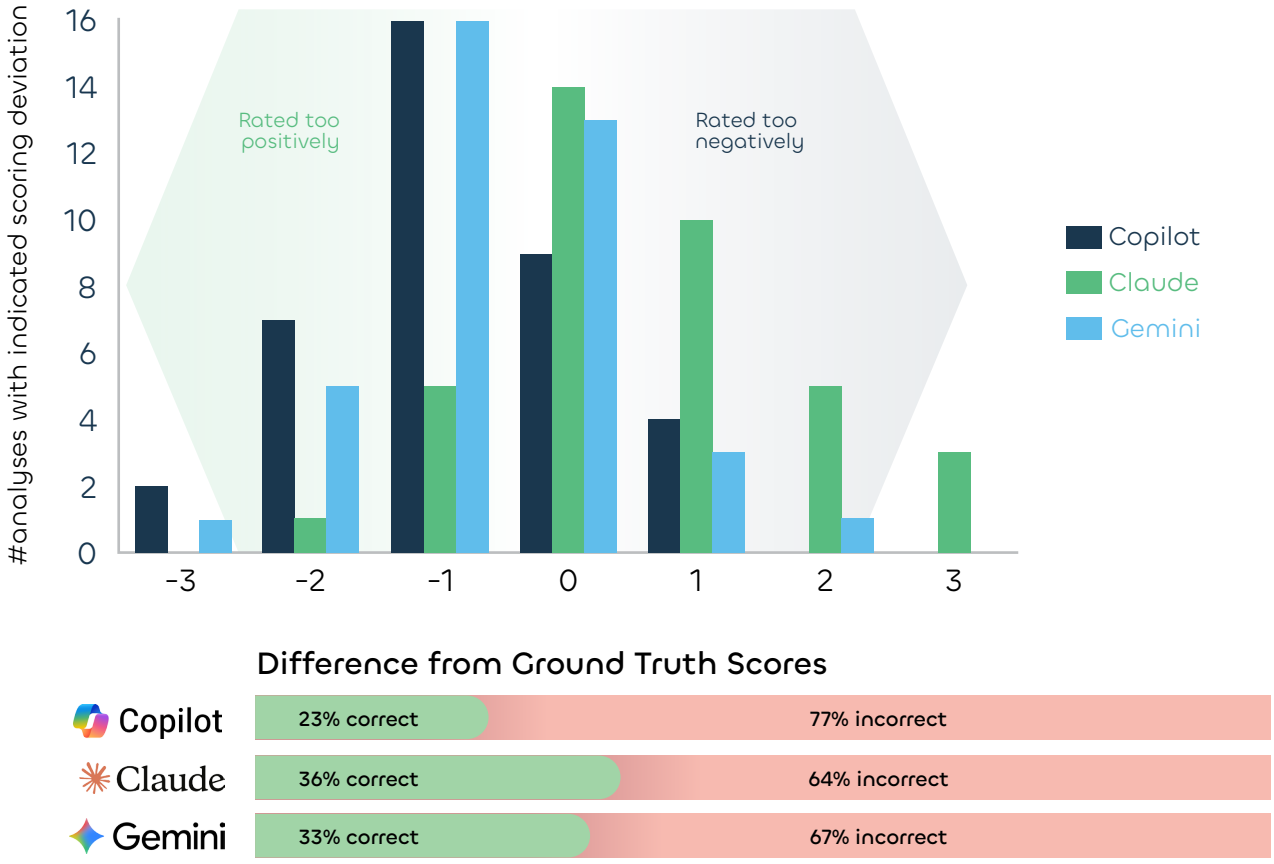


Figure 1. Benchmark results Bar plot showing the difference between model-assigned scores and Scitaris' ground truth. The x-axis represents Scitaris score minus the model score. Here, a negative score indicates an overly favorable rating of the model, while positive scores indicate an overly poor rating of the model. Overall, Copilot misjudged 77% of in-licensing opportunities, Claude misjudged 64%, and Gemini misjudged 67%.

Observed Limitations

Three systemic issues emerged:

- **Data Gaps** – Models frequently missed relevant information or relied on misinformation (e.g., incorrect MoA assignment leading to asset dismissal – systematic databases, company materials and expert analysis identified the asset as an ROCK/MYLK4 inhibitor, while Claude did not)
- **Hallucinations** – Models did hallucinate non-existing data. This had detrimental effects on subsequent analysis. For example, the model provided numerical response data from a poster source. However, upon closer inspection, the poster did not contain any of the claimed data.
- **Inferential Constraints** – Models lacked the capacity for reasoning beyond explicit input, often assigning scores despite insufficient data, or were unable to extrapolate. Models did not distinguish levels of data quality, which led, in cases, to overly optimistic judgments of assets based on merely qualitative data.

Implications

These findings underscore the inherent limitations of current LLMs in high-stakes decision-making without domain-specific augmentation.[1-3]

Specifically, the tendency of Copilot to lack critical or strongly positive scores prohibits any decision-making, while the random critical scores from Claude carry the risk of inducing wrong decisions.

At Scitaris, we address these gaps by:

- Integrating a curated knowledge base built after years of specialized consulting, e.g., indication-specific benchmarks and scientific requirements for success
- Embedding proprietary expert-driven frameworks into model workflows

This approach enables LLMs to deliver deeper, more reliable insights aligned with industry standards.

Across all evaluated assets, **Copilot misjudged 77% of in-licensing opportunities**, **Claude misjudged 64%**, and **Gemini misjudged 67%**, indicating that off-the-shelf LLMs are not yet reliable for high-stakes in-licensing decisions.

Human expertise remains essential for complex analysis

The primary objective of this study was to benchmark the capabilities of current LLMs in performing highly complex analytical tasks. Our findings reveal that approximately three-quarters of all initial model predictions required further refinement, underscoring both the potential and the limitations of these technologies.

While LLMs can provide valuable input for complex analyses, their outputs are frequently compromised by several challenges:

Current LLM performance challenges

Sycophancy bias
The model's tendency to generate responses designed to please rather than challenge.[4]

Hallucinations
Generation of plausible but factually incorrect information.[5]

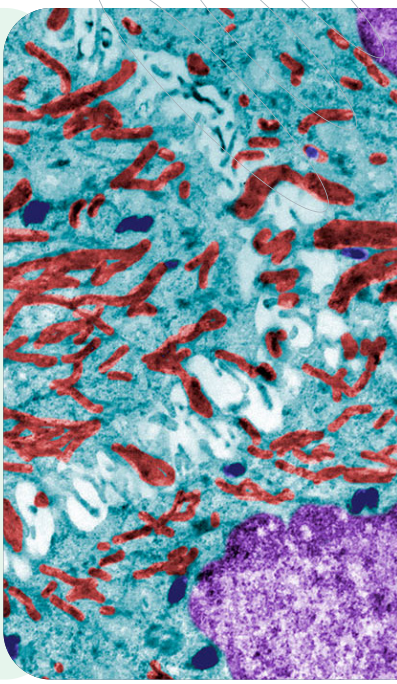
Data gaps
Incomplete or outdated source material and limited access to proprietary databases.[6]

Lacks critical evaluation
Equal treatment of all evidence regardless of quality or reliability.[6]

At Scitaris, we address these limitations through a systematic approach that combines:

1. Thoughtful integration of AI technologies into established workflows
2. Rigorous benchmarking against expert evaluation
3. Incorporation of the latest technical advancements
4. Continuous manual review and validation by domain experts

This hybrid methodology enables us to harness the efficiency and breadth of LLMs while maintaining the critical judgment and quality standards essential for high-stakes pharmaceutical business development decisions.[7] Moving forward, our commitment to this balanced approach will ensure that AI serves as a powerful augmentation tool rather than a replacement for human expertise.



Contact Us

About the Authors

Wignand Mühlhäuser, PhD

Senior Consultant

Wignand earned his B.Sc. in Applied Biology at Heinrich-Heine-University Düsseldorf, where he explored RNAi mechanisms and developed a strong interest in translational research. He completed his Master's at Albert-Ludwigs University Freiburg, focusing on synthetic biology and building his first optogenetic tools. During his PhD, he optimized and expanded optogenetic toolkits for controlling oncogenic signaling proteins and received the GBM Young Innovators Award.

Jonathan Vonnemann, PhD

Managing Partner

Jonathan is Managing Partner and co-founder of Scitaris, with deep expertise in biopharma consulting, scientific strategy, and global drug development. He earned his M.Sc. and Ph.D. summa cum laude in a joint program across five universities, including Cambridge, focusing on complex biological interactions. His research includes the first extension of the Cheng-Prusoff equation to multivalent interactions between biological surfaces such as viruses and cells.

References

1. Singhal, K. et al. (2025). Toward expert-level medical question answering with large language models. Nat Med. https://www.nature.com/articles/s41591-024-03423-7

2. Hager, P. et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. https://www.nature.com/articles/s41591-024-03097-1

3. Yu, E. et al. (2025). Large language models in medicine: applications, challenges, and future directions. Int J Med Sci. https://www.medsci.org/v22p2792.htm

4. Sharma, M. et al. (2024). Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.13548

5. Ji, Z. et al. (2023). Survey of hallucination in natural language generation. ACM Comput Surv. https://dl.acm.org/doi/10.1145/3571730

6. Ullah, E. et al. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology: a recent scoping review. Diagn Pathol. https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-024-01464-7

7. Aydin, S. et al. (2025). Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support. Front Med (Lausanne). https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1527864/full

SCITARIS
BIOPHARMA STRATEGY