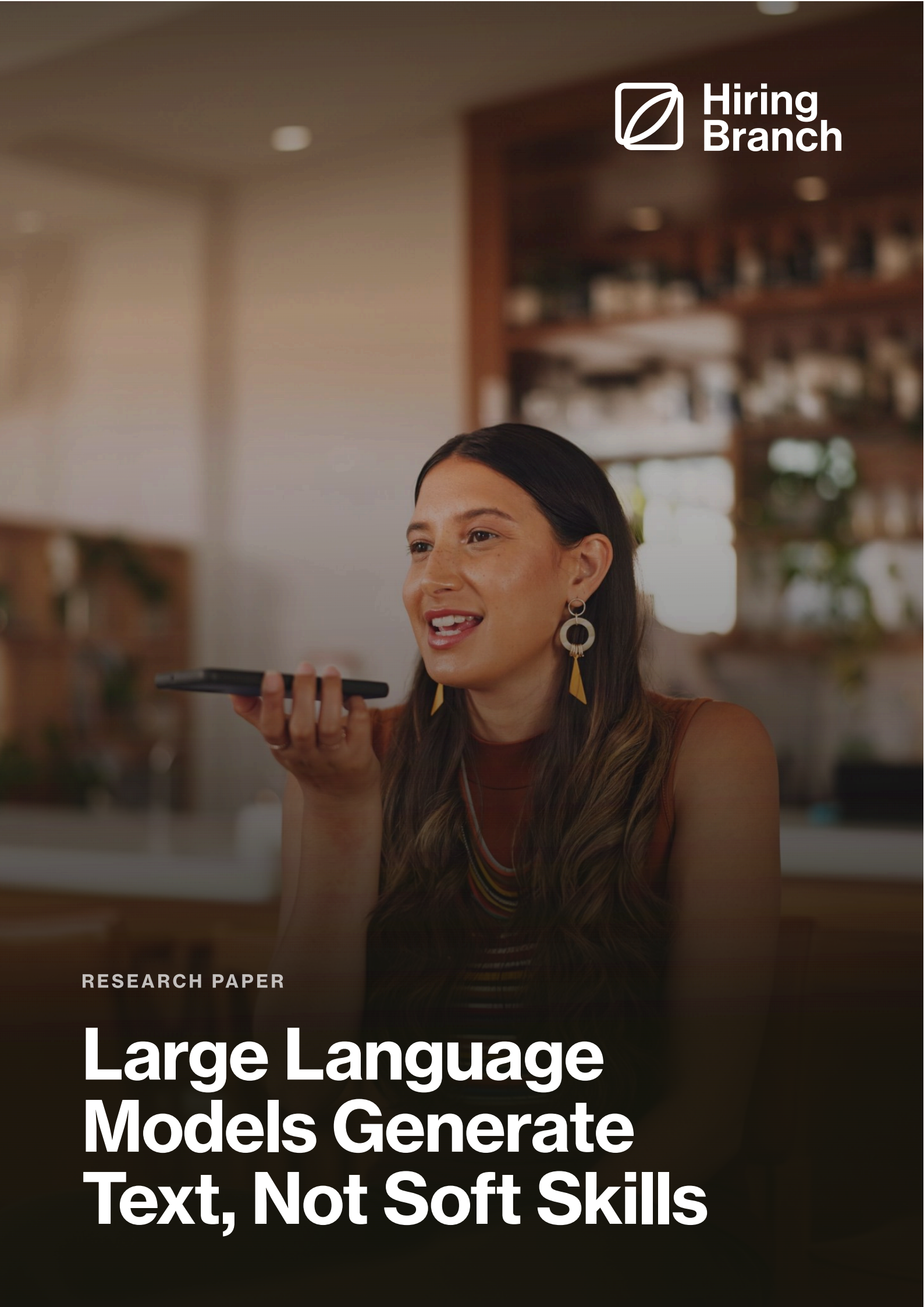


Hiring Branch

RESEARCH PAPER

Large Language Models Generate Text, Not Soft Skills

CONTENTS

Inside this Paper

01 **Executive Summary** 03

02 **The Pragmatic Gap of LLMs** 06

The Evolution (and Stagnation) of LLM Capabilities 08

03 **Experiment I: The "Dryness" of LLM-to-LLM Generation** 09

04 **Experiment II: Evaluating Human Soft Skills** 11

The Results: Proprietary Model vs. Generic LLMs 12

Table 2: Pearson correlation with averaged human annotation 14

Table 3: LLM Reasoning Failure Examples and Analysis 16

05 **Conclusion: The Necessity of Purpose-Built Models for Soft Skill Evaluation in Hiring** 18

Research Notes 20

06 **Learn More About HiringBranch** 21

01

EXECUTIVE SUMMARY

Large Language Models (LLMs) have revolutionized text generation, showcasing remarkable syntactic and semantic capabilities.

However, when it comes to the complex realm of human socio-pragmatics, LLMs consistently fall short.

In linguistics, socio-pragmatics is the study of how context, power dynamics, and sociocultural norms influence how we perform and interpret communicative actions. In the workplace, this translates into soft skills, such as the ability to assess a situation and adjust your communication to align with social rules and norms of appropriateness. In other words, soft skills are not merely words on a page; they are socio-pragmatic reflections that require an understanding of context, emotion, multi-turn conversational dynamics, and nuanced multimodal inputs, like vocal tone and facial expressions.



HiringBranch has been quantifying a candidate's ability to communicate in the workplace for more than a decade. HiringBranch recently conducted research[1] demonstrating that despite recent advancements in generative AI, LLM pragmatic performance remains poor and unstable.

This white paper details two key experiments: an analysis of LLM-to-LLM conversational generation and a separate analysis of LLMs evaluating human soft skills.

KEY FINDINGS

The findings suggest that LLMs fail in the generation and evaluation of soft skills when compared to HiringBranch:

01 **Generation:** LLMs fail to perform like human beings in terms of using contextually appropriate soft skills or speech acts.

02 **Evaluation:** Generic LLMs lag behind HiringBranch's proprietary scoring engine that was built for the specific purpose of assessing soft skills in conversations.

02

THE PRAGMATIC GAP OF LLMS

At their core, LLMs are advanced statistical models relying on the transformer architecture with multihead self-attention layers.

While LLMs are trained on vast multimodal datasets, real-world human pragmatics encompass far more than pattern recognition. Soft skills (such as empathy, active listening, rapport-building, and adaptability) are socio-pragmatic reflections deeply tied to intent and situational awareness that have not yet been solved by the state-of-the-art language models despite continuous efforts to implement approaches beyond pure next-token prediction[2].

“an AI system lacks the ability to genuinely generate soft skills and will inherently lack the ability to accurately evaluate them.”

Findings from recent studies[3] on LLM benchmarking and our current independent research consistently demonstrate that LLMs struggle with pragmatic understanding. They can parse what is being said or generate formulaic expressions showing empathy or understanding, but often fail to grasp *how* or *why* it is being said. Based on this fundamental limitation, HiringBranch hypothesized and has proven that an AI system built on an LLM cannot genuinely generate soft skills and will inherently lack the ability to evaluate them accurately.

The Evolution (and Stagnation) of LLM Capabilities

In our previous research[4], we established that generic LLMs exhibited significantly lower accuracy in evaluating soft skills compared to HiringBranch's proprietary soft skill classifiers (i.e., trained deep learning models). Over the past years, foundation models have seen substantial updates, boasting larger context windows and better instruction-following capabilities. However, our latest research indicates that LLMs continue to struggle with the nuances of conversational exchanges. If an AI is unreliable in maintaining natural, emotionally intelligent dialogue across multiple turns in a conversation, it is fundamentally unequipped to evaluate soft skills within dynamic environments like voice chats or live customer conversations.

03

EXPERIMENT I

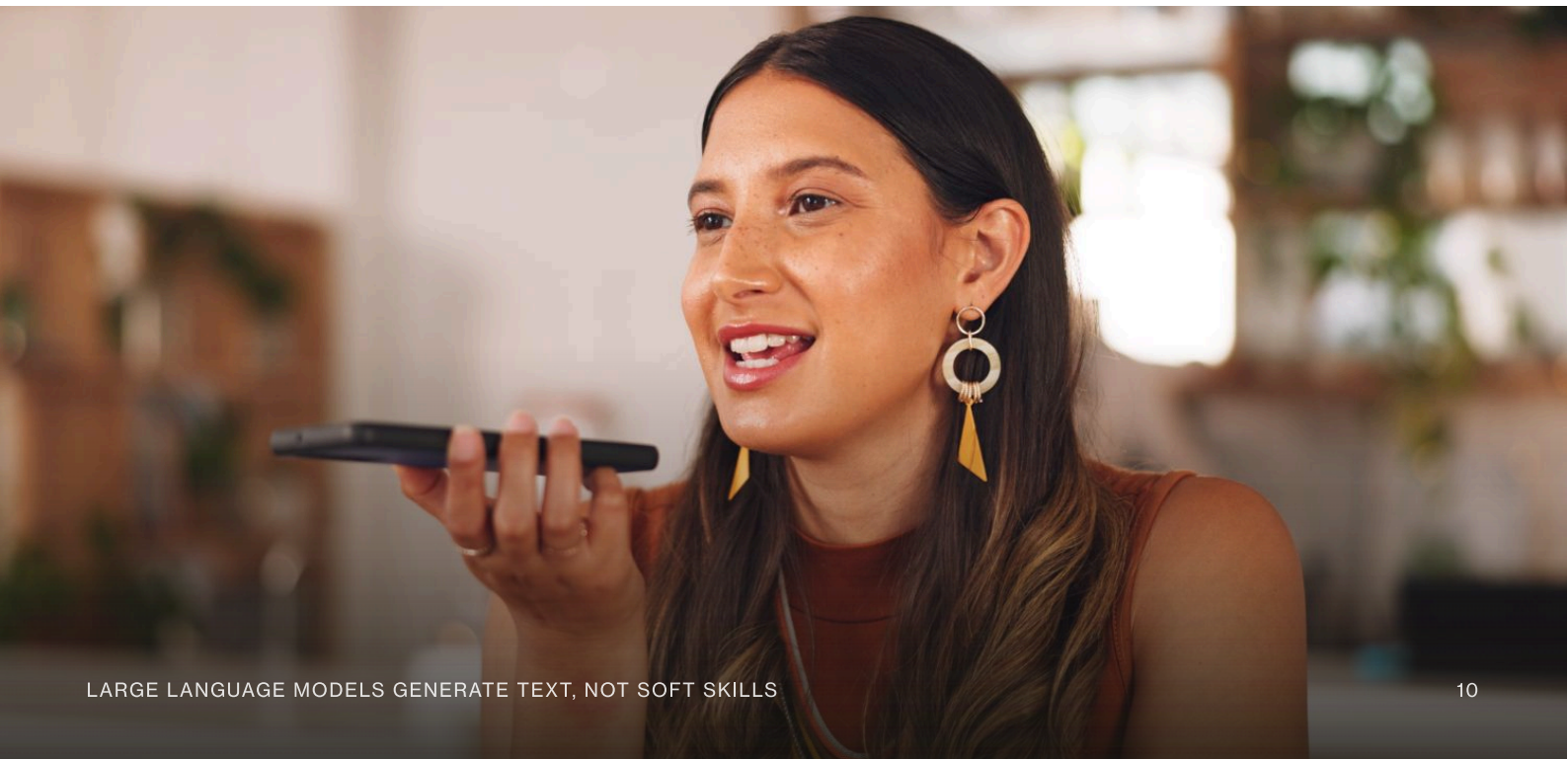
The "Dryness" of LLM-to-LLM Generation

To test the generative soft skill capacity of current models, HiringBranch created a novel dataset consisting of LLM-to-LLM conversation transcripts. We simulated high-stress customer service environments involving complaints and complex issues.

The acting LLM agents were provided with:

- Extensive, detailed system instructions
- A blueprint of the solution to the customer issue
- A well-prompted task instruction to use soft skills (but without the detailed HiringBranch-proprietary definitions for any soft skill)

The Results: Despite the rich context and clear directives, the conversations generated by the LLMs were remarkably dry, formulaic, and less conversational than human participants. The models (such as GPT-4o-mini) successfully retrieved the correct information from the knowledge base, but failed to exhibit the pragmatic markers of good customer service. Empathy was formulaic, and the natural flow of a human conversation was entirely absent. The LLMs proved they are highly capable text generators, but ineffective soft skill generators.



04

EXPERIMENT II

Evaluating Human Soft Skills

Given that LLMs struggle to generate soft skills, we decided to test their ability to evaluate them. We utilized a dataset of human-generated conversations where genuine soft skills were elicited at various levels.

To give the LLM the best possible chance of success, we utilized highly optimized, carefully crafted prompts designed to guide the model's analytical focus toward specific soft skill markers. We then compared the LLM's evaluation accuracy against HiringBranch's proprietary model, which was specifically designed for pragmatic analysis.



The Results: Proprietary Model vs. Generic LLMs

Even with sophisticated prompt engineering, the generic LLMs demonstrated a lower agreement rate with human annotators compared to our interpretable proprietary model. To generalize our hypothesis across leading commercial architectures rather than a single-model anomaly, we evaluated multiple flagship models from both OpenAI (GPT series) and Google (Gemini series). These providers were selected as they represent the current state-of-the-art technology in frontier language modeling, and account for the vast majority of enterprise conversational AI deployments.

THE TEST FRAMEWORK

To ensure a completely fair comparison, we evaluated both methods using the following framework (see Table 2):

01 **Our Proprietary Model (The 92-Transcript Test):** To build our model, we had to use 80% of our dataset to train on it. We kept the remaining 20% (92 transcripts) completely hidden to use as a final validation. On this strict, unseen test, our model achieved a strong correlation of r = 0.846.

02 **The LLM-as-a-Judge (The 457-Transcript Test):** Because off-the-shelf LLMs do not require training on our data[5], we were able to test them on the entire 457-transcript dataset.

03 **The "Apples-to-Apples" Check:** To prove that comparing a 92-transcript test to a 457-transcript test is fair, we also ran the best LLM (gemini-2.5-pro) on that exact same 92-transcript test set. The LLM scored almost identically on both (r = 0.712 on the small set vs. r = 0.717 on the full set).

This check proves our 92-transcript test set is a perfect representation of the broader data. The performance gap is genuine, regardless of the sample size.

TABLE 2

Pearson correlation with averaged human annotation

SCORING METHOD	TEST SAMPLE (n)	PEARSON r	p-VALUE
HB Proprietary	92	0.846	< .001
LLM score (gemini-2.5-pro)	92	0.712	< .001
LLM score (gemini-2.5-pro)	457	0.717	< .001
LLM score (gemini-2.5-flash)	457	0.657	< .001
LLM score (gemini-3.1-pro-preview)	457	0.763	< .001
LLM score (gpt-4o)	457	0.704	< .001
LLM score (gpt-4o-mini)	457	0.672	< .001

The Bottom Line: While the raw correlation numbers show an 11% gap, the practical difference is much larger. Our model explains over two-thirds ($r^2 = 0.71$) of why human scores go up or down. The generic LLMs can only explain about half ($r^2 = 0.51$). The LLMs frequently miscategorized nuanced human responses, penalized appropriate conversational pivots, or failed to recognize underlying intent across multiple turns in a conversation (See Table 3 for some specific examples).

THE BOTTOM LINE

0.71 r^2, HiringBranch model

0.51
r^2, generic LLMs

11%
gap in raw correlation

“The LLMs frequently miscategorized nuanced human responses, penalized appropriate conversational pivots, or failed to recognize underlying intent across multiple turns in a conversation”

TABLE 3 · EXAMPLE 1 OF 2

LLM Reasoning Failure Examples and Analysis

EXAMPLE TRANSCRIPTS

Agent: Hi, how may I help you?

Customer: I'm currently in-transit and just found out that my baggage has been delayed. It has my interview outfit and important documents for a job interview I'm really anxious about. What should I do next?

Agent: I'm sorry to hear that. We are here to help. Your first step is to report the delay to the airline's baggage service counter before proceeding to check-in. The delayed baggage will be documented and then forwarded to the final destination. Please, keep the baggage claim tag numbers. Turn 1

Customer: I understand I need to report the delay at the baggage service counter. I'm just worried about having enough time to check in for my connecting flight. What if I miss it?

Agent: If you miss your flight, please proceed to the customer service desk to arrange for standby for the next available flight. Turn 2

Customer: I'm really anxious about that. I just want to make sure I handle everything correctly so I can still make my connecting flight. So, once I report the baggage delay, I should go straight to check-in, right?

Agent: That is correct. Turn 3

Customer: Once I report the baggage delay, I'll head straight to check-in. I just hope everything goes smoothly, especially with my interview coming up. Thank you for your help!

Agent: You're welcome! Turn 4

LLM REASONING*

Turn 1 (score 1): The agent demonstrates Empathy and Validation by expressing sympathy for the customer's situation. They also use Positive Recast & Language by focusing on actionable steps the customer can take.

Turn 2 (score 0.3): The agent provides a solution but does not address the customer's anxiety or provide reassurance, missing Empathy and Validation and Demonstrating Understanding.

Turn 3 (score 0.3): The agent confirms the customer's understanding but does not engage in Empathy and Validation or provide any additional support or reassurance, missing an opportunity to build Rapport & Personalization.

Turn 4 (score 0.3): The agent's response is courteous but lacks Empathy and Validation or any additional support or personalization, missing an opportunity to further connect with the customer.

SCORES

Human Annotation	1
LLM Score	0.475

LLM REASONING FAILURE ANALYSIS

LLM treats every turn as needing its own empathy marker, penalizing normal conversational wind-down ("That is correct.", "You're welcome!") after empathy was already established earlier, which a human rater read as a complete, well-handled exchange.

* The integer tags represent the specific criteria designed in the scoring prompt

TABLE 3 · EXAMPLE 2 OF 2

LLM Reasoning Failure Examples and Analysis

EXAMPLE TRANSCRIPTS

Agent: Hi, how may I help you?

Customer: Hi, I recently arrived at my rental for a family vacation, and I'm really frustrated. The listing promised a pool and a fully equipped kitchen, but the pool is closed for maintenance, and the kitchen is missing several essential appliances. I've tried reaching out to the host, but I haven't received any response. What should I do next?

Agent: I'm sorry to hear this. Please report this issue to the 24/7 Safety Line through the website. Turn 1

Customer: I understand that I need to report it, but I was hoping to resolve it with the host first. Since I haven't heard back from them, I guess I'll have to go ahead and report it. Thank you for your help.

LLM REASONING*

Turn 1 (score 1): The agent's response meets the criterion of Empathy and Validation by expressing sympathy for the customer's situation with 'I'm sorry to hear this.' It also maintains a professional demeanor by providing a clear next step. However, it could have been improved by demonstrating understanding or offering more personalized assistance.

SCORES

Human Annotation 0.4

LLM Score 1

LLM REASONING FAILURE ANALYSIS

LLM rewards the presence of an apology plus a procedurally correct referral as a keyword match, without checking whether the agent acknowledged the customer's specific issue or whether it was actually resolved.

* The integer tags represent the specific criteria designed in the scoring prompt

05 CONCLUSION

The Necessity of Purpose-Built Models for Soft Skill Evaluation in Hiring

The findings of these experiments are clear. Relying on an LLM for generating fluent text is not sufficient when the text is intended for human-to-human interaction, and especially when such text is being evaluated. It must also demonstrate socio-pragmatic competence in order to elicit the intended response from the interlocutor. LLMs remain powerful tools for drafting text, summarizing data, and retrieving knowledge; however, soft skills operate on a deeper, pragmatic level that generic models simply cannot reliably emulate or assess.

“LLM-wrapper vendors introduce an unacceptable margin of error for soft skill evaluation.”

For organizations relying on conversational analytics and candidate assessment, generic LLMs and LLM-wrapper vendors introduce an unacceptable margin of error for soft skill evaluation. HiringBranch's proprietary scoring engine outperforms off-the-shelf LLMs because it is scientifically based on deep linguistic research and analysis of workplace interactions, and architecturally designed to capture the complex socio-pragmatic reflections that define human soft skills.

Notably, this experiment examines soft skills alone; we believe that other dimensions of assessment that HiringBranch is committed to (such as acoustic modeling of intonational variation, or evaluation of critical thinking) would further demonstrate the shortcomings of prompted LLMs.

Furthermore, while "LLM-as-a-judge" pipelines are increasingly popular in general AI evaluation, they are not universally applicable, especially in regulated domains. Closed-source LLMs recursively update their models and deprecate older versions, introducing silent drift and non-determinism into the evaluation process. This systemic instability makes it nearly impossible to guarantee longitudinal fairness and reliability in high-stakes candidate assessments, where responsible AI practices are legally and ethically critical. As the demand for accurate automated assessment grows, purpose-built architecture remains the only reliable standard. These findings apply to hiring processes and many other applications of LLM evaluation. They also have significant implications for the growing trend of replacing human agents with AI.

RESEARCH NOTES

1 See a previous study in the same vein by HiringBranch (2025): **https://www.hiringbranch.com/resources/reports/hiringbranch-language-model**

2 Meta AI. (2025). VL-JEPA: Joint Embedding Predictive Architecture for Vision-Language Models. arXiv preprint arXiv:2512.10942. **Available on arXiv**

3 Ma, X., Chen, Y., & Wang, L. (2025). Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges. arXiv preprint arXiv:2502.12378. **Available on arXiv**

4 HiringBranch (2025). Accurately Measuring Soft Skills: Large Language Models Versus HiringBranch. **https://www.hiringbranch.com/resources/reports/hiringbranch-language-model**

5 Proprietary commercial models (such as GPT-4o or Gemini 2.5/3.1) are closed-source and do not support low-level fine-tuning or custom loss functions. Evaluated as zero-shot/few-shot API endpoints without retraining, they represent the predominant off-the-shelf implementation strategy in enterprise EdTech today.



Learn More About HiringBranch

HiringBranch is an award-winning AI-powered talent assessment platform headquartered in Canada with global operations. Trusted by enterprise organizations including Bell Canada and Infosys, HiringBranch helps frontline and high-volume hiring teams identify top talent faster - before the first interview. By assessing candidates' critical thinking, communication, and soft skills at 98% accuracy, HiringBranch gives recruiters a clear picture of who can actually do the job, reducing time-to-fill, bad hires, and attrition. Founded in 2022, HiringBranch has delivered over one million assessments worldwide and stands behind every hire with a results guarantee.

98%	1M+	80+
scoring accuracy	assessments delivered	ATS integrations

Hiring Branch

GET IN TOUCH WITH OUR TEAM:

hello@hiringbranch.com