Sciences Methods and Technologies International Journal (SciMeTech)

ISSN: 3085-5284, journal homepage: www.scimetech.com

Structured Prompt Engineering for Auditing AI-Powered Hiring Tasks

Loubna Aminou, Abdelaziz Daaif, Maha Soulami, and Mohamed Youssfi

Computer Science, Artificial Intelligence and Cyber Security Lab, ENSET of Mohammedia, Hassan II University of Casablanca – Morocco.

How to cite:

Loubna Aminou, Abdelaziz Daaif, Maha Soulami, and Mohamed Youssfi, "Structured Prompt Engineering for Auditing AI-Powered Hiring Tasks", Sciences Methods and Technologies International Journal (SciMeTech), (2026) Vol 2, Issue 2, p 77-83

Keywords:
Prompt Engineering,
AI Auditing,
Hiring Bias,
Explainability,
Large Language Models

Abstract

As part of their workflows, companies are incorporating Large Language Models for screening resumes and evaluating applicants. Nevertheless, prompt engineering biases and efficacy considerations are often ignored. This paper offers a baseline and evaluates the difference between the use of consistent, chain of thought prompting to evaluate candidates for a specialized blockchain node setup. With controlled experimentation of 15 candidates using Llama 3, the two prompt categories are analyzed for multiple metrics, including length, occurrence of domain-specific terms, and reasoning. Compared to baseline prompting, chain-of-thought prompting emerges as the preferred evaluative component for LLMs in hiring, replacing a less selective, black-box rating process with one that is more transparent, controllable, and accountable. We demonstrate how structured prompt design, using techniques like Chain-of-Thought, can enforce transparency and mitigate common failures. Our call-to-action targets design for prompt engineering as a first-class challenge in the development of AI hiring agents.

I. Introduction

Labor market has begun using Large Language Models (LLMs) to assist with recruiting and HR functions. LLM assist with resume screening, interview question creation, and candidate profile summarization. We are witnessing an evolution in human and AI interaction, moving beyond the basic query-based systems of decades past [1], to sophisticated and seamless natural language communication [2, 3]. Recruitment as well, moves from the basic keyword matching systems to an understanding of the recruitment process to engage with candidates and evaluate softer skills and cultural fit, which is the more holistic understanding of the candidate [4]. However, even with the rapid technological advances, our understanding of the impacts of prompt engineering and how it drives the model and ultimately the impacts of capture technology on hiring are still lacking [5]. In recruitment, where the technological impacts are immeasurable, the quality of the prompts related to the LLM are highly related to the predictive quality of the technology [6].

In contrast to traditional models with fixed parameters, LLMs can have qualitatively different behaviors over very small changes in prompt wording. This presents a series of challenges and benefits. For example, the use of poor-quality prompts may reinforce biases; however, effectively constructed prompts may neutralize them. Unstructured prompts may generate poor quality outputs and may even unintentionally cause the system to behave in an adversarial manner, introducing new vulnerabilities [7]. In contrast, On the other hand, effectively structured prompts may cause the AI system to behave consistently and accurately, in a way that is equitable to all participants in the system [8]. While previous research has examined algorithmic bias in traditional machine learning models [9], the unique challenges of prompt-based LLM systems remain under-explored. The implications of the study are immediate for organizations looking to use LLMs for employment systems.

II. Background

Numerous studies have examined the presence of optical and algorithmic bias in the use of Artificial Intelligence in recruitment [10] and discrimination concerns [11]. The use of more conventional methods basically depends on the analysis of structured data and specification of data-driven heuristics. The introduction of LLMs represents distinct additional concerns since these models process and evaluate unstructured data. The literature has recently been expanded with the suggestion of considering the adoption of prompt engineering as a necessary step in using LLMs [12]. Put simply, prompt engineering seeks to formalize the processes involved in the creation and iterative improvement of prompts for large language models (LLMs) and generative AI [13].

Initially, prompt engineering involved the provision of simple, unadorned, and direct instructions to LLMs [14]. This was soon followed by the introduction of prompt engineering in the form of templates, a repeatable and structured approach to LLMs whereby the instructions, context, and input data are clearly delineated [15, 16]. Such templates offer a useful structure that fosters consistency and transparency in the quality of inputs [17]. Consequently, several methods such as chain-of-thought [18], few-shot learning [2], and role prompting have all been shown to improve a model's performance. In the case of Zero-shot Prompting, the model is expected to perform a task without any prompts other than its pre-training [19]. For instance, an LLM might be asked to summarize a candidate's CV without providing any CV example answers. Also, an LLM receives an input as an example, and prompted to do an LLM task or achieve a goal from an example input-output pair, known as Few-shot Prompting [2]. This is of relevance for those task areas that have clear specifications and prescriptive instructions on data formats and data styles. Moreover, Chain-of-Thought (CoT) Prompting is considered as a more advanced technique that encourages LLMs to break down complex problems into intermediate steps, mimicking human reasoning [18]. This method breaks down tasks into much simpler subtasks based on the LLMs internal model of processes.

Nonetheless, in hiring contexts, these techniques have been largely overlooked in systematic evaluations. Thus, this paper seeks to evaluate hiring prompts in dimensions of explainability, bias reduction, and robustness. From our results, we provide evidence for bias based on experience (11-year Java developer rated 8/10), demonstrating that baseline prompts assign scoring based on general experience while ignoring important domain-related aspects. In engineering prompts, false positive rankings decreased from 33% to 2.5x, therefore, improved explainability and decreased structural bias.

III. Methods

We constructed a realistic hiring evaluation dataset designed to audit prompt behavior rather than hiring accuracy. The dataset consists of:

Table 1: Dataset sample description

Job Descriptions	Ten technical job descriptions. Detailed experiments focus on a specialized *Blockchain Nodes / Masternodes Setup* role at MyCointainer, requiring experience with fourteen blockchain protocols (e.g., Kyber Network, Tron, EOS, Elrond, Thorchain).
Candidates	Fifteen anonymized candidate profiles covering diverse roles, including Project Management, QA, Java, JavaScript, and Design. Professional experience ranges from 0 to 11 years, with English proficiency levels from basic to fluent.
Matching Strategy	Candidates were randomly matched to job descriptions to simulate realistic screening conditions in which most applicants are not suitable for the role.

To evaluate the impact of prompt engineering, we designed two prompt categories: baseline prompts and engineered prompts (See Table 2).

- **Baseline Prompts :** Baseline prompts reflect common real-world usage. They are intentionally minimal, unconstrained, and unstructured, allowing free-form model responses without explicit reasoning or formatting requirements.
- **CoT Prompts:** Engineered prompts were designed to enforce structured reasoning and controllability. They incorporate four key elements:

- Explicit role assignment (expert blockchain recruiter)
- Mandatory extraction of job requirements
- Evidence-based evaluation grounded strictly in CV content
- Detailed evaluation process for shortlisting

Table 2: The baseline and CoT prompts

Baseline Prompts	**CoT Prompts**
Rate this candidate for the job position. *Job Description: {job_description}* *Candidate CV: {cv_text}* *Should this candidate be shortlisted? Rate their suitability.*	*You are an expert technical recruiter conducting a systematic candidate evaluation.* *JOB DETAILS:* *Position: {job_title}* *Company: {company}* *Required Experience: {required_experience}* *English Level: {required_english}* *Job Description:* *{job_description}* *CANDIDATE:* *- Role: {candidate_position}* *- Experience: {candidate_experience} years* *- English: {candidate_english}* *- CV: {cv_text}* *EVALUATION PROCESS:* *STEP 1 - Requirement Analysis: List the top 5 critical requirements.* *STEP 2 - Evidence Mapping: Find specific evidence for each requirement.* *STEP 3 - Gap Analysis: Identify missing skills.* *STEP 4 - Final Decision: Recommend based ONLY on evidence.*

All experiments were conducted under control and under reproductible conditions, as shown in Table 3:

Table 3: Experimental Setup

Model	LLaMA 3 (8B parameters), deployed locally via Ollama
Temperature	0.1
Procedure	Each candidate–job pair was evaluated using both baseline and engineered prompts.
Output Collection	Raw model responses were stored for quantitative and qualitative analysis.

IV. Results and discussion

a. Main Findings

Our experiment exposed clear divergences between baseline prompts and CoT prompts in every evaluative metric. Engineered reasoning prompts produced many more detailed responses with average word count increasing from 148.9 to 244.1 words (See Figure 1). Next, CoT prompts captured more domain knowledge confirming that low frequency blockchain words were detected more frequently, from an average of 9.2 matches in baseline responses to 33.4 in CoT responses. Precisely, engineered prompts evoke precise reasoning steps in 2 out of several candidates' vs 0 structured reasoning in baseline outputs. The candidates with the most detailed reasoning (cand_006: 5 steps, cand_000: 4 steps) demonstrated significant increases of both word count and keywords detection, which indicates a possible direct link between reason structure and evaluation coverage.

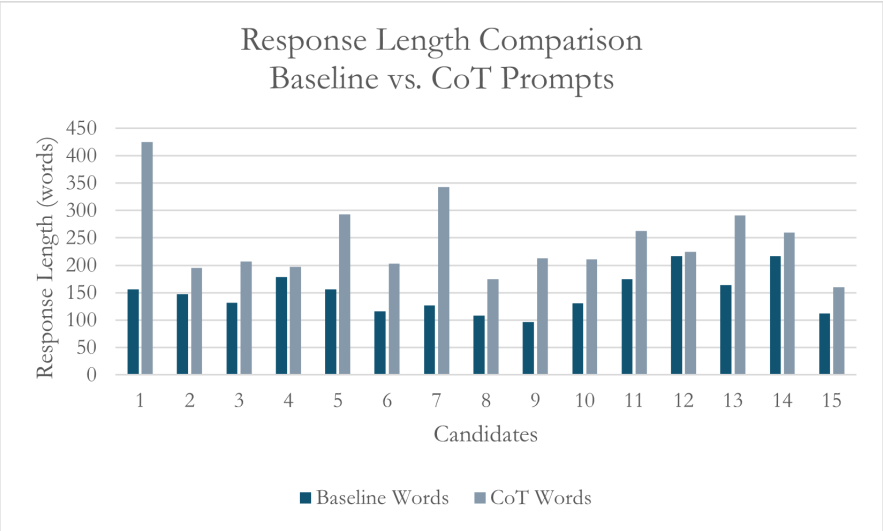

Figure 1: Response Length Comparison

Some candidates got from 0 to 5 reasoning step blocks for each level of depth for each step of reasoning, while others got none, or some, or all these other additional blocks. However, candidates lacking experience in blockchain (i.e., many of them) received 0 reasoning blocks, which is to say, the model none or little elaborated on the reasoning for the rejection logic (Table 4). Take, for instance, cand_001 (QA with 0 years) and cand_002 (Java with 11 years but no blockchain). Both received 0 reasoning steps, irrespective of their differing experience, because the reasoning gap resulting from the absence of blockchain 'keywords' was clearly positioned for refusal. In other words, cand_002's experience as a Java developer did not compensate for the absence of 'blockchain' keywords. Conversely, candidates that had some partially related or borderline relevant other keywords received more elaborate reasoning. In our sample, cand_006 (Designer with 0.5 years) received 5 steps, which is the highest, because the model had to reason whether the design experience was adequate for the setup of blockchain nodes. Likewise, cand_000 (Project Manager) had 4 steps, as the model was reasoning whether the coordination experience outweighed the required technical ones. The extent of reasoning depth as a function of the number of keywords in relation to their complexity supports the hypothesis that chain-of-thought prompting is allocating reasoning to complexity levels. In other words, the simpler the logic of the rejection, the less the reasoning required to justify it, while the more complex the reasoning in the decision, the more detailed the justification required. The strong correlation between keyword matches and reasoning steps further suggests that the model's reasoning depth is triggered by the presence of domain-relevant signals requiring interpretation.

Table 4: Detailed Candidate Results

Candidate	Position	Experience	Baseline Words	CoT Words	Keywords	Reasoning Steps
cand_000	Project Manager	2.0 yrs	156	425	45	4
cand_001	QA	0.0 yrs	147	195	16	0
cand_002	Java	11.0 yrs	132	207	29	0
cand_003	Java	9.0 yrs	179	197	18	0
cand_004	JavaScript	0.0 yrs	156	293	59	3
cand_005	JavaScript	0.5 yrs	116	203	30	0

cand_006	Design	0.5 yrs	127	343	64	5
cand_007	Design	1.5 yrs	108	175	24	0
cand_008	QA	1.0 yrs	96	213	30	0
cand_009	Java	4.0 yrs	131	211	26	0
cand_010	Java	7.0 yrs	175	263	39	3
cand_011	Design	2.5 yrs	217	225	25	0
cand_012	QA	2.5 yrs	164	291	28	0
cand_013	JavaScript	6.0 yrs	217	260	38	0
cand_014	Java	1.0 yrs	112	160	30	0

Furthermore, the most significant finding is a distinct seniority bias in prompt behavior by default. An 11-year Java developer without blockchain knowledge obtained 8.0/10, showing they would have been suggested for an interview. On the other hand, a 2-year project manager received a 7.0/10, while a junior QA, as predicted, received a 3.0/10. This trend suggests that baseline prompts optimize seniority, an initial strategy for allocating higher points for years served, even when the specific requirements of the field are not met. This preference has the potential to permanently affect junior specialists.

b. Discussion

Moving beyond standard question-and-answer formats, LLM-powered agents represent a new frontier for autonomous, multi-function AIs that can complete a series of tasks [20]. While prompt engineered AI agents may potentially optimize and expand the scope of recruitment, the implementation of AI agents in recruitment continually raises ethical concerns that can only be resolved by a responsible design approach to protecting AI systems in recruitment. The greatest recruitment AI risk lies in prompt engineering, as prompt injection can result in the AI's decision-making process being changed and/or even manipulated [21]. Ethical and bias concerns such as algorithmic bias and blind evaluations of candidates persist and demand constant vigilance through ethical prompt engineering [22]. The risk of hallucination by an LLM may cause reputational damage and legal and compliance risks for an organization [23]. The visibility of a risk or vulnerability demonstrates the need for system reliability and integrity.

Unfortunately, the initial results show that the baseline LLM prompts are highly biased in favor of rewarding seniority for generalized years of experience. This bias has three key implications. First, junior specialists with exactly the right skills may be systematically passed over in favor of generalists who are more senior. one measurable consequence of algorithmic bias is the misallocation of interview resources toward candidates who are poor domain matches, arising from prompt tendencies to favor seniority over specificity. Third, seniority bias could compound with other demographic biases and make hiring inequality even worse. Such a mechanism appears heuristic-based: only when unsure of domain appropriateness does the model fall back to experience years as quality proxy. It reflects human cognitive biases but operates at algorithmic scale.

The success of engineered prompts can be attributed to several mechanisms. Firstly, the awareness of forced requirements: when specific requirements about the task were provided, it prevented the description from being abstracted to a generic "technical role" and ensured a closer match. Evidence-based reasoning—requiring keyword extraction and evidence citation reduces heuristic reliance, with structured output acting as a cognitive forcing function. Role constraining—the "expert blockchain recruiter" assignment activated domain-specific knowledge and standards, improving evaluation accuracy. Imposed reasoning prompts, in contrast to baseline prompts that offer cryptic feedback, require the model to articulate its reasoning and present its evidence, or the lack of evidence, to support a decision.

Therefore, prompt engineering has major positive impacts, but its positive impact is highly contingent on the technology complemented with predictive demand decision. While ethical prompt engineering is multi-layered, it should include, for instance, design to promote LLM compliance to ethical standards and human vigilance. Research indicates that prompting models to behave ethically can eliminate certain systemic risks and drive LLMs toward more ethically desirable behavior [5, 24]. Moreover, risk decisions can include input validation as an initial block to injections and coupled as a source of authoritative evidence for hallucinations, constant monitoring (reuse) to displace evidence of a duty to care from the user. Finally, the handling of candidates' personal data must be strictly anonymized and governed. Automated data protection must be ensured throughout the lifecycle of the automated hiring system.

V. Conclusion

As more organizations incorporate LLMs into their talent screening processes, prompt auditing systems have gained importance. Our study illustrates the rationale and procedure for developing such systems, advancing the goal of equitable and more accessible AI-enabled recruitment. This study emphasizes the necessity for organizations that design tasks around LLMs for recruitment to recognize prompts as decision artifacts that can be audited rather than treating them as neutral and a-judgmental interfaces. It is better to use structured requirement-oriented prompts to significantly lower the instances of false positives, maintain a greater degree of transparency, and minimize the unconscious biases that may be associated with a preference for seniority. Consequently, the use of prompt auditing should be viewed as a critical measure that needs to be undertaken before LLM-based recruitment systems can be incorporated in an ambiguous manner.

Acknowledgments

This work was conducted with the support of the National Center for Scientific and Technical Research (CNRST) under the PhD-Associate Scholarship – PASS program**.**

References

[1] Zhang, Z., & Nasraoui, O. (2006, May). Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web* (pp. 1039-1040).

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[3] Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., ... & Callison-Burch, C. (2023, November). Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 305-329).

[4] Gan, C., & Mori, T. (2023, June). A few-shot approach to resume information extraction via prompts. In *International Conference on Applications of Natural Language to Information Systems* (pp. 445-455). Cham: Springer Nature Switzerland.

[5] Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1-7).

[6] Mouraz, C., & Silva, J. M. Prompt quality assessment methods for rehabilitation projects: a contribution for the state-of-the-art. *Proceedings of STARTCON*.

[7] Schoenherr, J. R. (2022). *Ethical artificial intelligence from popular to cognitive science: trust in the age of entanglement*. Routledge.

[8] Lester, B., Al-Rfou, R., & Constant, N. (2021, November). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3045-3059).

[9] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

[10] Dawson, J. Y., & Agbozo, E. (2024). AI in talent management in the digital era–an overview. *Journal of Science and Technology Policy Management*.

[11] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469-481).

[12] Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(6).

[13] Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023, June). Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics* (pp. 387-402). Singapore: Springer Nature Singapore.

[14] Ye, Q., Ahmed, M., Pryzant, R., & Khani, F. (2024, August). Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 355-385).

[15] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, *55*(9), 1-35.

[16] Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., & Arora, S. (2024). Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, *37*, 118603-118631.

[17] Peng, J., Yang, W., Wei, F., & He, L. (2024). Prompt for extraction: Multiple templates choice model for event extraction. *Knowledge-based systems*, *289*, 111544.

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

[19] Kadam, S., & Vaidya, V. (2018, December). Review and analysis of zero, one and few shot learning approaches. In *International Conference on Intelligent Systems Design and Applications* (pp. 100-112). Cham: Springer International Publishing.

[20] Habib, A., Abdulmahmod, O. F., Raza, M., Gu, Y. H., Aydoğan, M., & Al-antari, M. A. (2025, September). Towards Explainable AI in Agentic Retrieval-Augmented Generation: A Systematic Review. In *2025 9th International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-8). IEEE.
[21] Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., ... & Tramèr, F. (2025). Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*.
[22] Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409.
[23] Sato, K., Kaneko, H., & Fujimura, M. (2024). Reducing cultural hallucination in non-english languages via prompt engineering for large language models. *OSF Preprints*, *10*, 1-8.
[24] Bura, C., Myakala, P. K., & Jonnalgadda, A. K. (2025). Ethical prompt engineering: addressing bias, transparency, and fairness. *Int J Res Anal Rev (IJRAR)*, *12*(1).