Human Systems Engineering and Design (IHSED2025), Vol. 198, 2025, 337–342
https://doi.org/10.54941/ahfe1006787

LLMs as Retail Cart Assistants: A Prompt-Based Evaluation

Ratomir Karlović and Ivan Lorencin

Faculty of Informatics, Juraj Dobrila University of Pula, Zagrebačka 30, 52100 Pula, Croatia

ABSTRACT

Large language models (LLMs) offer promising capabilities for interpreting user input in natural language and translating it into structured formats for downstream processing. This study investigates the use of LLMs as shopping-cart assistants, limited to the task of parsing natural-language commands into a predefined JSON schema containing three fields: action, product, and quantity. The objective is to evaluate the models' ability to perform accurate semantic parsing under consistent conditions. To examine the impact of prompt design, three distinct prompting strategies were developed: a minimal instruction specifying the target fields, an extended prompt including synonym guidance and formatting rules; and a few-shot learning approach incorporating multiple examples with strict output requirements. Each prompt variant was applied identically across all selected LLMs to ensure comparability. The evaluation was conducted using a dataset of 1,000 synthetic shopping-cart commands generated via a large generative AI model. Each command was paired with a known ground truth, structured into the same target schema. Model-generated outputs were transformed into CSV format and compared against these references to assess parsing performance. By systematically varying prompt complexity and controlling for model input, this study provides a controlled comparison framework for assessing prompt effectiveness in narrow, structured tasks. The results contribute to a deeper understanding of prompt design as a determinant of LLM utility in applied, goal-oriented scenarios.

Keywords: LLM, Retail smart assistant, Prompt engineering

INTRODUCTION

The rapid evolution of retail platforms has necessitated the development of sophisticated tools to enhance the customer shopping experience. A critical component of this experience is the ability of users to efficiently manage their shopping carts. Traditionally, shopping assistants have faced challenges such as task-specificity and poor generalization, requiring dedicated models for various functionalities and struggling with new product integration (Zhang et al., 2024).

Large Language Models (LLMs) have emerged as powerful tools, revolutionizing various applications through their advanced natural language processing capabilities and superior generalization (Karn, 2025; Wei et al., 2025). Their ability to understand and generate human-like text

makes them highly suitable for developing omnipotent assistants capable of handling diverse tasks (Zhang et al., 2024). However, while LLMs demonstrate immense potential in various domains, including scientific data extraction (Lee et al., 2024) and educational assessment (Khan et al., 2025), their specific performance as shopping cart assistants—translating natural language requests into structured actions—remains an area requiring focused investigation. This is particularly true when considering the impact of different prompt engineering strategies on their accuracy and efficiency (Rubei et al., 2025).

This study addresses this research gap by systematically comparing the performance of 11 diverse LLMs, encompassing both proprietary OpenAI models and open-source Ollama models, in acting as a shopping cart assistant. This comparative analysis aims to provide valuable insights into the efficacy of various LLMs and prompt engineering techniques for enhancing user interaction.

METHODOLOGY

This study evaluated the performance of 11 large language models (LLMs) in parsing natural-language shopping-cart commands into structured JSON objects. Each model was tasked with transforming user requests into a JSON format containing three fields: action (either "add" or "remove"), product (as mentioned in the request), and quantity (an integer, defaulting to 1 if unspecified).

Three prompt variants were designed and uniformly applied across all models. The first prompt provided minimal instruction, simply requesting extraction of the required fields. The second prompt included additional guidance on acceptable synonyms and formatting conventions. The third prompt employed a few-shot learning strategy, providing multiple structured examples and stricter format constraints. No prompt was fine-tuned or customized for individual models.

The evaluated models include open-source models run via Ollama, covering a range of parameter sizes and architectures. The full set of tested models was as follows:

Open-source models run locally via Ollama

- ollama/llama3.3:70b-instruct-q2_K,
- ollama/llama3.2:3b,
- ollama/llama3.1:8b,
- ollama/qwen3:8b,
- ollama/qwen3:4b,
- ollama/deepseek-r1:14b,
- ollama/deepseek-r1:1.5b,
- ollama/mistral:7b,
- ollama/phi4:14b,
- ollama/gemma3:4b,
- ollama/granite3.3:8b.

The evaluation dataset consisted of 1,000 synthetic user commands generated using a generative AI model. Each command was paired with a

manually verified ground truth JSON label specifying the intended action, product, and quantity. Each model was prompted independently with the same set of 1,000 commands under each of the three prompt conditions. Model outputs were parsed into CSV format and compared row-wise against the ground truth.

RESULTS

The performance evaluation of eleven large language models (LLMs) revealed distinct differences in parsing accuracy across models and prompt types. Using a row-wise F1 score as the primary evaluation metric, all 1,000 user commands were processed by each model under three prompting conditions: minimal instruction, extended guidance, and few-shot learning. The results indicate a consistent trend of performance improvement as prompts became more detailed, though the extent of improvement varied by model.

Under the minimal instruction prompt, the best-performing model was Deepseek-r1_14b with an F1 score of 0.991, followed by Phi4_14b (0.974), Qwen3_8b (0.969), and Qwen3_4b (0.962). Larger models such as LLaMA3_3_70B-instruct-q2_K and Granite3_3_8b also performed well with F1 scores of 0.956. Models like Mistral_7b (0.947), LLaMA3_2_3b (0.939), and Gemma3_4b (0.937) achieved competitive results, while LLaMA3_1_8b lagged slightly behind at 0.907. The lowest performance was observed in Deepseek-r1_1_5b, which achieved an F1 score of 0.534. The average F1 score across all models with minimal prompts was 0.916 (see Figure 1), indicating modest baseline parsing performance in the absence of detailed guidance.

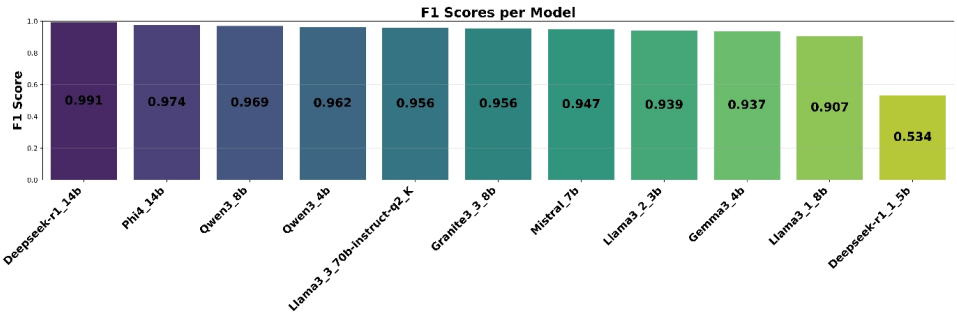

Figure 1: F1 score for minimal instruction prompt.

With the extended prompt, which included synonym handling and explicit formatting instructions, broad performance gains were observed. Deepseek-r1_14b maintained its lead with a near-perfect score of 0.997, followed closely by LLaMA3_3_70B-instruct-q2_K (0.988), Qwen3_8b (0.985), and Qwen3_4b (0.982). Several other models also showed strong results: Granite3_3_8b (0.973), LLaMA3_1_8b (0.969), Mistral_7b (0.969), and Phi4_14b (0.958). Notably, Gemma3_4b and LLaMA3_2_3b performed well with scores of 0.958 and 0.932, respectively. Deepseek-r1_1_5b improved

slightly to 0.607, but remained the weakest model. The average F1 score under extended prompting rose to 0.938 (see Figure 2), confirming the benefit of explicit prompt design.

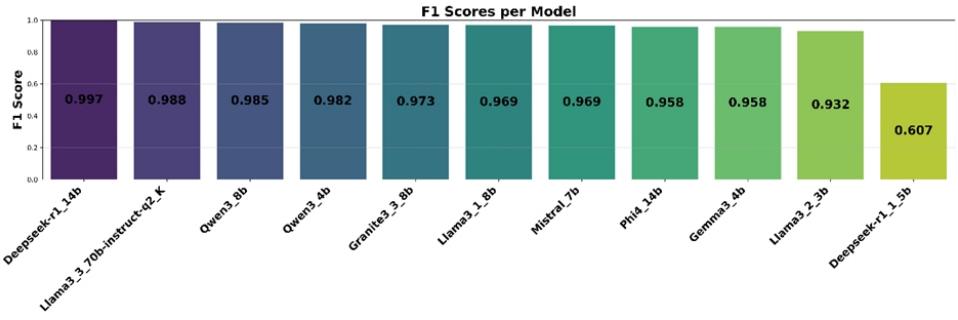

Figure 2: F1 score for extended prompt.

The few-shot prompting condition, which provided multiple structured examples, yielded the highest overall scores for top-tier models. LLaMA3_3_70B-instruct-q2_K achieved a perfect F1 score of 1.000, with other high performers including Phi4_14b (0.994), LLaMA3_1_8b (0.990), Qwen3_4b (0.988), and Deepseek-r1_14b (0.979). Smaller models such as Gemma3_4b (0.970) and LLaMA3_2_3b (0.958) also exhibited substantial improvements. However, not all models benefited equally: Qwen3_8b (0.948), Granite3_3_8b (0.944), Mistral_7b (0.827), and Deepseek-r1_1_5b (0.598) showed more limited performance gains. The average F1 score across all models in the few-shot condition was 0.927 (see Figure 3), slightly lower than the extended prompt average but with a higher density of near-perfect scores among the top-performing models.

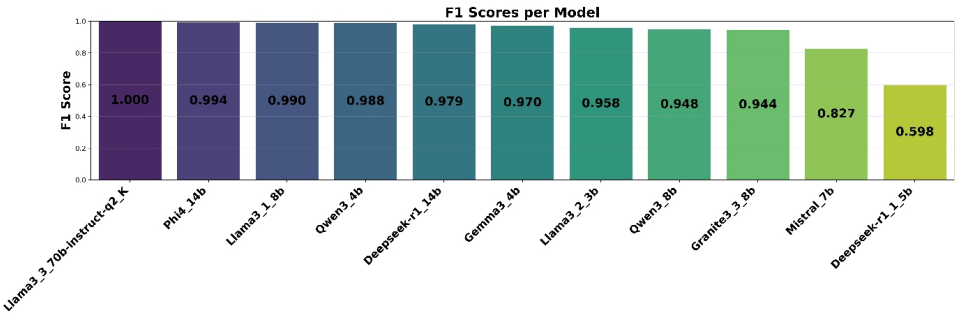

Figure 3: F1 score for few-shot prompt.

Across all prompt types, Deepseek-r1_14b and LLaMA3_3_70B-instruct-q2_K consistently ranked among the best performers, showing strong robustness to prompt variation. In contrast, Deepseek-r1_1_5b consistently underperformed, highlighting the influence of model scale and tuning on task performance. Interestingly, while few-shot prompting produced the

highest individual scores, extended prompting achieved a higher average score, suggesting that structured verbal instruction yields more consistent gains across different architectures.

These findings emphasize the importance of both model selection and prompt design in optimizing LLM performance for structured information extraction. While increasing prompt complexity improves accuracy, the benefits are not uniform, reflecting a nuanced relationship between model architecture and prompt formulation.

CONCLUSION

This study systematically evaluated the parsing capabilities of eleven large language models on a structured shopping-cart command task, highlighting the significant role of prompt design in model performance. The results show that both model architecture and prompt complexity—ranging from minimal instructions to fully structured few-shot examples—substantially influence parsing accuracy. Few-shot prompting produced the highest F1 scores for several models, including a perfect score by LLaMA3_3_70B-instruct-q2_K, reinforcing the value of concrete examples in guiding model behavior.

However, extended prompts led to the highest average performance across all models, suggesting that clearly structured instruction can be more universally effective. Models such as Deepseek-r1_14b and LLaMA3_3_70B-instruct-q2_K emerged as top performers under all conditions, demonstrating strong adaptability. In contrast, Deepseek-r1_1_5b consistently underperformed, likely due to its limited capacity or tuning.

These findings emphasize the importance of prompt engineering as a critical tool for enhancing LLM performance in domain-specific tasks. The demonstrated performance improvements from minimal to extended and few-shot prompts suggest practical strategies for deploying LLMs in real-world applications requiring accurate structured data extraction. Future work should explore dynamic prompt generation, fine-tuning approaches, and cross-lingual generalization to further improve outcomes and expand applicability.

Overall, this research offers a detailed benchmark and practical insights for leveraging LLMs in natural language command parsing, contributing to their effective integration in structured interaction scenarios.

ACKNOWLEDGMENT

This research is (partly) supported by SPIN projects “INFOBIP Konverzacijski Order Management (IP.1.1.03.0120)”, “Projektiranje i razvoj nove generacije laboratorijskog informacijskog sustava (iLIS)” (IP.1.1.03.0158), “Istraživanje i razvoj inovativnog sustava preporuka za napredno gostoprimstvo u turizmu (InnovateStay)” (IP.1.1.03.0039), “European Digital Innovation Hub Adriatic Croatia (EDIH Adria) (project no. 101083838)” under the European Commission’s Digital Europe

Programme and the FIPU project “Sustav za modeliranje i provedbu poslovnih procesa u heterogenom i decentraliziranom ra cunalnom sustavu”.

REFERENCES

Karn, R. R. (2025). Linear Feedback Control Systems for Iterative Prompt Optimization in Large Language Models. arXiv preprint arXiv:2501.11979.

Khan, M. A., Ayub, U., Naqvi, S. A. A., Khakwani, K. Z. R., Sipra, Z. B. R., Raina, A.,... & Riaz, I. B. (2025). Collaborative large language models for automated data extraction in living systematic reviews. Journal of the American Medical Informatics Association, ocae325.

Lee, W., Kang, Y., Bae, T., & Kim, J. (2024). Harnessing large language model to collect and analyze metal-organic framework property dataset. arXiv preprint arXiv:2404.13053.

Rubei, R., Moussaid, A., Di Sipio, C., & Di Ruscio, D. (2025). Prompt engineering and its implications on the energy consumption of Large Language Models. arXiv preprint arXiv:2501.05899.

Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W. (2025). Research on Intelligent Grading of Physics Problems Based on Large Language Models. Education Sciences, 15(2), 116.

Zhang, S., Peng, B., Zhao, X., Hu, B., Zhu, Y., Zeng, Y., & Hu, X. (2024). LLaSA: Large language and e-commerce shopping assistant. arXiv preprint arXiv:2408.02006.