Revista Innovación y Software
Vol. 6, No. 2, Mes Septiembre-Febrero, 2025
ISSN: 2708-0935
Pág. 6-11
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos cortos
Temática: Ingeniería de software
Recibido: 10/8/2024 | Aceptado: 16/9/2024 | Publicado: 30/9/2025

Identificadores persistentes:
DOI: 10.48168/innosoft.s24.a212
ARK: ark:/42411/s24.a212
PURL: 42411/s24.a212

Optimización de Modelos de Lenguaje Grande (LLMs) a través del Prompt Engineering

Optimization of Large Language Models (LLMs) through Prompt Engineering

Sergio Díaz Sifuentes[1][0009-0001-7550-5922]*, **Crishtian Paz Fernández**[2][0009-0000-2027-5990], **Marcelino Torres Villanueva**[3][1234-1111-2222-3333]

[1]Universidad Nacional de Trujillo. sdiazsi@unitru.edu.pe
[2]Universidad Nacional de Trujillo. t1023300121@unitru.edu.pe
[3]Universidad Nacional de Trujillo. mtorres@unitru.edu.pe

*Autor para correspondencia: sdiazsi@unitru.edu.pe

Resumen
Este artículo exploró el impacto del prompt engineering en la optimización del rendimiento de modelos de lenguaje grande (LLMs, por sus siglas en inglés) como GPT y BERT. El prompt engineering fue presentado como un enfoque innovador que consistía en diseñar instrucciones específicas para guiar las respuestas de los modelos, mejorando su precisión y relevancia sin modificar sus parámetros internos. El estudio evaluó metodologías para la construcción de prompts efectivos, comparó diferentes estrategias como el few-shot y el zero-shot learning, y analizó casos prácticos en áreas como la generación de texto, la respuesta a preguntas y el análisis de sentimientos. Los resultados mostraron que un diseño estratégico de prompts podía mejorar significativamente la calidad de las respuestas, reducir errores y ampliar el rango de aplicaciones de los LLMs.

Palabras claves: Few-shot learning, LLMs, modelos generativos, prompt engineering, zero-shot learning.

Abstract
This article explored the impact of prompt engineering on optimizing the performance of large language models (LLMs) such as GPT and BERT. Prompt engineering was introduced as an innovative approach that involved designing specific instructions to guide the models'responses, enhancing their accuracy and relevance without modifying their internal parameters. The study evaluated methodologies for constructing effective prompts, compared different strategies such as few-shot and zero-shot learning, and analyzed practical cases in areas like text generation, question answering, and sentiment analysis. The results demonstrated that a strategic design of prompts could significantly improve response quality, reduce errors, and expand the range of LLM applications.

Keywords: *Few-shot learning, generative models, LLMs, prompt engineering, zero-shot learning*

Introducción

La evolución de los Modelos de Lenguaje Grande (LLMs, por sus siglas en inglés) revolucionó el procesamiento del lenguaje natural (NLP), proporcionando capacidades avanzadas para comprender, analizar y generar texto con una fluidez similar a la humana. Estos modelos, como GPT (Generative Pre-trained Transformer) y BERT (Bidirectional Encoder Representations from Transformers), establecieron nuevos estándares en aplicaciones como asistentes virtuales, generación de resúmenes, traducción automática y análisis semántico. Su éxito se basó en las arquitecturas de Transformadores, introducidas por Vaswani et al. (2017), que utilizan mecanismos de atención para procesar grandes volúmenes de datos lingüísticos de manera eficiente [3].

A pesar de sus destacadas capacidades, los LLMs enfrentaron desafíos en su implementación práctica. Uno de los principales problemas fue que su desempeño dependía significativamente de cómo se les presentaba cada tarea, incluyendo el diseño de instrucciones que guiaran al modelo hacia respuestas precisas y relevantes. En este contexto, el prompt engineering emergió como una técnica innovadora que permitía optimizar el uso de estos modelos sin necesidad de ajustes computacionales extensivos.

El prompt engineering se fundamentó en la formulación de instrucciones específicas que maximizaban el rendimiento del modelo en tareas particulares. Estrategias como el zero-shot learning, que se basaba en instrucciones claras sin ejemplos previos, y el few-shot learning, que empleaba ejemplos dentro del prompt para proporcionar contexto, demostraron ser efectivas para resolver problemas complejos (Brown et al., 2020) [1]. Además, técnicas como el razonamiento en cadena (chain-of-thought prompting) permitieron descomponer problemas en pasos secuenciales, mejorando la precisión del modelo en tareas lógicas y matemáticas (Wei et al., 2022) [6].

Este estudio se enfocó en explorar el impacto del prompt engineering como una herramienta para optimizar las capacidades de los LLMs. Analizó cómo esta técnica no solo incrementaba la precisión y relevancia de las respuestas, sino que también facilitaba su adopción en aplicaciones prácticas, desde la atención al cliente hasta la investigación científica. El objetivo fue establecer una base metodológica que permitiera maximizar el potencial de los modelos existentes mientras se abordaban sus limitaciones actuales, como la generalización y la sensibilidad a instrucciones ambiguas.

Materiales y métodos o Metodología computacional

La investigación sobre la optimización de Modelos de Lenguaje Grande (LLMs) a través del prompt engineering se realizó empleando un enfoque metodológico integral. Este combinó la revisión sistemática de literatura, análisis de casos prácticos, experimentos controlados y entrevistas a expertos en la materia. Estas metodologías,

utilizadas en conjunto, permitieron explorar el impacto y las posibilidades de esta técnica en el rendimiento de los LLMs, así como identificar áreas clave para futuras investigaciones. A continuación, se detallan los enfoques metodológicos utilizados:

1. **Revisión de la literatura** Se llevó a cabo una exhaustiva revisión de la literatura científica relacionada con los LLMs y el prompt engineering. La revisión abarcó publicaciones académicas de alto impacto, artículos técnicos, informes de investigación y documentación oficial de desarrolladores como OpenAI y Google. Este análisis permitió identificar conceptos clave, estrategias utilizadas y las limitaciones reportadas en la implementación de prompts para optimizar tareas específicas en modelos como GPT-4 y BERT [1, 2, 5].

2. **Análisis de casos de éxito** Se seleccionaron casos prácticos de organizaciones y equipos de investigación que han implementado con éxito técnicas de prompt engineering. Los casos incluyeron aplicaciones en generación de texto técnico, análisis de sentimientos y respuestas automatizadas. Se analizaron desafíos enfrentados, soluciones propuestas y métricas de desempeño logradas, destacando cómo estrategias como few-shot learning y razonamiento en cadena han permitido mejoras significativas en el rendimiento de los LLMs [1, 6].

3. **Metodologías experimentales** Se realizaron experimentos controlados para evaluar la eficacia de diferentes estrategias de prompt engineering. Los experimentos incluyeron tareas de generación de texto, clasificación temática y resolución de problemas lógicos. Se compararon los enfoques zero-shot, few-shot y razonamiento en cadena en términos de precisión, coherencia y relevancia de las respuestas [6].

   Herramientas utilizadas: modelos preentrenados como GPT-4 y BERT, implementados mediante bibliotecas como Transformers de Hugging Face [4]. Métricas de evaluación: F1-score, precisión y métricas personalizadas relacionadas con satisfacción del usuario.

4. **Entrevistas a expertos Metodologías experimentales** Se realizaron entrevistas estructuradas con investigadores, desarrolladores y profesionales que utilizan LLMs en aplicaciones prácticas. Estas entrevistas proporcionaron información valiosa sobre las tendencias emergentes en prompt engineering y los desafíos técnicos enfrentados durante su implementación. También ayudaron a identificar oportunidades de mejora en el diseño y uso de prompts.

5. **Análisis integrado** Finalmente, se llevó a cabo un análisis integrado de los hallazgos obtenidos a través de las metodologías mencionadas. Este análisis permitió identificar patrones comunes, validar resultados y generar conclusiones fundamentadas. También se destacó la relación sinérgica entre las estrategias experimentales y los hallazgos de la revisión de literatura y las entrevistas, estableciendo una base sólida para recomendaciones futuras [3, 7].

Resultados y discusión

Los resultados de la investigación sobre la optimización de los Modelos de Lenguaje Grande (LLMs) mediante prompt engineering han demostrado que esta técnica mejora significativamente el rendimiento en tareas específicas de procesamiento del lenguaje natural. Al diseñar instrucciones detalladas y contextuales, se logra una mayor precisión y relevancia en las respuestas de los modelos, especialmente en aplicaciones como la generación de texto, la clasificación temática y la resolución de problemas complejos.

Los experimentos realizados muestran que el prompt engineering puede reducir la dependencia de ajustes computacionales complejos, lo que facilita la implementación práctica de LLMs en diversas industrias. Además, se destacó que esta técnica mejora la adaptabilidad de los modelos a diferentes contextos y dominios.

En general, los resultados confirman que el prompt engineering es una herramienta crucial para maximizar el potencial de los LLMs, ofreciendo soluciones rápidas, precisas y adaptables a las necesidades específicas de los usuarios.

A continuación, se presentan los resultados obtenidos de la investigación:

1. **Automatización de Tareas Complejas** El prompt engineering permite a los LLMs automatizar tareas avanzadas como la generación de informes, análisis de datos y clasificación de textos. Por ejemplo, en experimentos con GPT-4, los prompts optimizados generaron resúmenes técnicos con una precisión del 85 %, superando significativamente las expectativas de tareas automatizadas previas [1, 7].

2. **Mejora en la Precisión de las Respuestas** La capacidad de diseñar prompts contextuales incrementó la precisión en la resolución de problemas matemáticos y en la clasificación temática. Los modelos que utilizaron razonamiento en cadena (chain-of-thought prompting) lograron aumentar la precisión en un 20 % en comparación con los enfoques tradicionales de few-shot learning [6].

3. **Reducción de Falsos Positivos** Al utilizar estrategias de prompt engineering, se observó una disminución significativa en los errores de clasificación y falsos positivos en tareas de análisis de sentimientos. Esto optimizó el uso de recursos y mejoró la eficiencia en aplicaciones prácticas como la atención al cliente [1, 4].

4. **Personalización de Respuestas** El uso de prompts adaptados a contextos específicos permitió personalizar las respuestas de los LLMs según las necesidades de los usuarios. Por ejemplo, en el análisis de datos financieros, los prompts diseñados con datos históricos relevantes mejoraron la utilidad de las respuestas en un 30 % [3, 5].

5. **Predicción de Respuestas Complejas** Los modelos preentrenados, como GPT-4, mostraron una notable capacidad para predecir y estructurar respuestas en tareas de alta complejidad al analizar patrones previos y datos históricos proporcionados en los prompts. Este enfoque permitió a los usuarios obtener respuestas más completas y útiles en menos tiempo [1, 7].

Conclusiones

El prompt engineering se consolida como una técnica esencial que está redefiniendo el uso de los Modelos de Lenguaje Grande (LLMs) en el ámbito del procesamiento del lenguaje natural. A pesar de los desafíos asociados a su implementación, como la dependencia de la calidad de los datos y la falta de interpretabilidad, su capacidad para optimizar la precisión, relevancia y adaptabilidad de los modelos demuestra un enorme potencial para aplicaciones prácticas en diversos dominios.

El éxito del prompt engineering radica en su habilidad para maximizar las capacidades de los LLMs sin requerir ajustes computacionales extensivos. Su capacidad para personalizar respuestas, reducir errores y mejorar la eficiencia en tareas específicas lo posiciona como una herramienta poderosa en entornos donde la precisión y la flexibilidad son críticas. Sin embargo, su adopción requiere un enfoque estratégico que integre tanto consideraciones técnicas como éticas, asegurando la transparencia y la responsabilidad en todas las etapas del proceso.

El desarrollo de talento especializado en prompt engineering es fundamental para garantizar su implementación efectiva. Esto no solo incluye habilidades técnicas para diseñar prompts efectivos, sino también una comprensión profunda de las implicaciones éticas y sociales de esta tecnología. La formación de profesionales en este campo será clave para abordar los desafíos asociados y capitalizar plenamente sus beneficios.

En resumen, el prompt engineering ofrece una oportunidad única para expandir las capacidades de los LLMs, pero su implementación debe estar respaldada por una planificación consciente, inversión en recursos humanos y una visión ética que asegure su impacto positivo en la sociedad. Esto permitirá no solo mejorar la eficiencia y efectividad de los modelos, sino también garantizar su alineación con valores y principios responsables en un mundo cada vez más impulsado por la inteligencia artificial.

Contribución de Autoría

Crishtian Brenon Paz Fernández: Conceptualización, Investigación, Metodología, Software, Validación, Redacción - borrador original.**Sergio Helí Diáz Sifuentes:** Conceptualización, Investigación, Metodología, Análisis formal, Recursos, Visualización, Supervisión, Administración de proyectos, Adquisición de fondos,

Curación de datos, Escritura, revisión y edición. **Marcelino Torres Villanueva.**

Referencias

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, ..., and D. Amodei, "Language models are few-shot learners," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-art natural language processing," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/1910.03771

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020. [Online]. Available: https://arxiv.org/abs/1910.10683

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2201.11903

[7] OpenAI, "GPT-4 technical report," OpenAI, 2023. [Online]. Available: https://openai.com/research/gpt-4