В. Ільїн

Науковий керівник – к.т.н., доц. Бачинський Р.В.

Віртуальний асистент пошуку інформації в документах на базі штучного інтелекту.

Проект присвячений розробці системи, здатної швидко і точно надавати відповіді на запити, ґрунтуючись на інформації, що міститься у користувацьких документах. Основна мета роботи полягає у створенні інтелектуального інструменту, який інтегрує пошук інформації з генерацією відповідей за допомогою техніки "пошуково доповненої генерації" (RAG). Це дозволяє ефективно використовувати наявні текстові дані для швидкого надання відповідей, що ґрунтуються на релевантних текстових уривках з документів.

Суть розробки полягає у тому, що віртуальний асистент обробляє запити користувачів, знаходить у базі документів відповідний абзац або уривок, який може містити відповідь, і передає його до моделі OpenAI. Модель OpenAI, на основі цього тексту, генерує відповідь, що допомагає точно і зрозуміло інтерпретувати запит, використовуючи контекст з документів. Наприклад, якщо користувач запитує про історію України, система знаходить відповідний абзац у документі, відправляє його до OpenAI, і генерується відповідь на основі даних з цього уривка. Таким чином, система не тільки здійснює пошук, а й підсилює його через генерацію контекстуалізованої відповіді, що є її ключовою перевагою.

Практичне застосування створеної системи полягає у використанні її в різних галузях, де необхідно швидко шукати інформацію в великих масивах тексту, таких як бізнес, юриспруденція, академічна діяльність та інші. Особливо корисною вона буде для професіоналів, яким потрібен доступ до спеціалізованих або конфіденційних даних, що зберігаються у приватних документах.

Основні переваги системи включають швидкість, точність і гнучкість у пошуку. На відміну від класичних пошукових систем або стандартних моделей генерації, віртуальний асистент на базі RAG поєднує контекстуальний пошук з можливістю вдосконалення відповіді. Це дозволяє не тільки знаходити релевантні частини документів, а й надавати логічно пов'язані відповіді, які враховують специфіку запиту. Крім того, створена система має власну пам’ять, що є значною перевагою у порівнянні з OpenAI, яка не забезпечує зберігання

інформації між запитами. Це дає можливість користувачам працювати з довготривалими запитами, отримувати більш точні та персоналізовані відповіді, а також зберігати важливі контексти між сеансами роботи.

Таким чином, розробка є своєрідною обгорткою для OpenAI, збагаченою власною пам'яттю та спеціальними можливостями, що забезпечує користувачам ефективний інструмент для пошуку та аналізу інформації в документах, який перевершує існуючі рішення у сфері автоматизації пошуку та обробки тексту.

Література:

1) *Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.*
2) *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 4171-4186).*
3) *Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (pp. 9459-9474).*
4) *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS) (pp. 5998-6008).*
5) *Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Pearson.*
6) *Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS) (pp. 3111-3119).*
7) *Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, 10(1), 1-309.*
8) *Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.*
9) *Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1(8), 9.*