Synthetic Data Generation with Large Language Models for Personalized Community Question Answering

Anonymous ACL submission

Abstract

Personalization in Information Retrieval (IR) is a topic studied by the community for a long time. However, the collection and curation of high-quality training data requires significant costs and time investment, especially for collecting user-related information. In this paper we explore the usefulness of Large Language Models (LLMs) in generating synthetic documents tailored to user's personal interests using user-related information. We introduce a new dataset, Sy-SE-PQA, to study the effectiveness of models fine-tuned on LLM-generated data and study how the complexity of personalization impacts model performances. We build Sy-SE-PQA based on an existing dataset, SE-PQA, which consists of questions and answers posted on the popular StackExchange communities. Starting from questions in SE-PQA, we generate synthetic answers using different prompt techniques and LLMs. Our findings suggest that LLMs have high potential in generating training data, tailored to user's needs, for neural retrieval models and it can be used to replace training data. The code is publicly available[1].

1 Introduction

Instruction fine-tuned Large Language Models (LLMs), such as GPT (Brown et al., 2020), have shown high capabilities in generating synthetic data for different tasks such as Relation Extraction in the medical domain (Tang et al., 2023), Dialogue Systems (Hämäläinen et al., 2023) and Information Retrieval (IR) (Bonifacio et al., 2022; Askari et al., 2023). We focus our attention on an Information Retrieval task, specifically Personalized Information Retrieval (PIR) . The topic of Personalized Information Retrieval has been studied by researchers for a long time (Speretta and Gauch, 2005; Borisov et al., 2016; Wu et al., 2017). The goal of PIR is to produce an output tailored to a specific user or group of users by leveraging their interests and online behaviour, which requires gathering user-related information so as to capture the user's interests and preferences. Existing datasets like the AOL query log (Pass et al., 2006), the Yandex query log [2]and the CIKM Cup 2016 [3] dataset, even if commonly used, have privacy concerns and limitations due to anonymization (Barbaro et al., 2006). The performance of PIR models significantly depends on the quality of the training data; this is a substantial challenge due to the scarcity of large-scale, publicly available datasets that include detailed user information, which hinders the training of effective personalized neural models. Recently, a new dataset, called SE-PQA (Kasela et al., 2024), has contributed to fill this gap; it is built by collecting questions and answers from StackExchange, the well-known community Question Answering (cQA) platform. The goal of SE-PQA is to favour the design of personalized Question Answering (QA) systems adapted to an Information Retrieval task, where the question is seen as a query, and the answers are retrieved from the pool of answers. SE-PQA contains questions (i.e. queries) from 50 different communities, which can be categorized under the large umbrella of humanistic communities.

Recent studies focus on generating synthetic data for IR tasks, where LLMs can be used to generate synthetic queries (Bonifacio et al., 2022) or documents (Askari et al., 2023). Wang et al. (2023) examined creating documents from queries and appending them to the queries for query expansion. Additionally, recent studies on ChatGPT have primarily explored its applications in ranking and retrieval tasks. Guo et al. (2023) created a dataset, called HC3, to compare human and ChatGPT answers to questions from four different domains:

[1] `https://anonymous.4open.science/r/SY_SE-PQA`

[2] Yandex Query Log

[3] CIKM Cup 2016

medicine (He et al., 2020), finance (Maia et al., 2018), Wikipedia (Yang et al., 2015) and Reddit (Fan et al., 2019). Faggioli et al. (2023) examined the potential of LLMs to generate relevance labels and Sun et al. (2023) evaluated ChatGPT's effectiveness in re-ranking candidate documents based on given queries. However, little effort has been devoted to generating documents tailored to users' needs, which we will call personalized documents. In this paper, to fill this gap, we analyze the ability of recent LLMs, namely, GPT-3.5 (Brown et al., 2020) and Phi-3 (Abdin et al., 2024), to generate an answer to a given question using the information related to the user who wrote the question. We train neural retrieval models with both synthetic and human-written data, analysing the advantages and limitations related to the use of LLMs for generating training data for a personalized IR task.

We discover that, even though the generated training data may not be factually accurate, the models trained with synthetic data outperform statistically significantly baseline models such as BM25 evaluated on a test set with human written answers, showing that training an IR model with factually true answers does not seem mandatory, although this hypothesis should be further enforced.

2 Methodology

In this section, we begin by outlining the process of generating synthetic data for our PIR task (Section 2.1). We then detail the models used to evaluate the effectiveness of our generated data (Section 2.2).

2.1 Dataset Generation

Our Sy-SE-PQA is based on the personalized version of the SE-PQA dataset (Kasela et al., 2024), which contains over 200k training questions sampled from 50 different StackExchange communities. Each question consists of a Title, summarizing the question, and a Body, providing context. Furthermore, when users ask a question, they can assign some tags, selected from a predefined set, to specify the topic and make the question easier to find for other users interested in it. We generate three different answers for each question, varying the levels of personalization and context information provided to the LLM:

1. **Basic Answer Generation:** For each question, we generate synthetic answers starting from the following prompt:
   "Write an answer to the given question: Title: [TITLE] Body: [BODY]."
2. **Personalized Answer Generation:** To generate personalized answers, we provide the five most frequently used tags by the user prior to the input question. We add the following to the initial prompt:
   "Answering the question, consider that who asks the question is interested in: [TAGS]. Ignore the user interests if they are not relevant to the question without mentioning that you have ignored them."
   This ensures that the model focuses on relevant interests only.
3. **Contextual Answer Generation:** To generate answers with contextual information about the community in which the question was written. The prompt is as follows:
   "Write an answer to the given question in the context of [COMMUNITY]: Title: [TITLE] Body: [BODY]."

In the last prompt, the contextual information regarding the community a question belongs to can be useful to answer it using domain knowledge. For example, the question reported in Table 5 in Appendix A: a user asks who is depicted in the statue featured in a specific scene from the movie `Ocean's 12`. The user prefers to write the question in the History community instead of the Movie one, probably, because he is interested in the historical aspect of the answer.

We rely on the following three LLMs to generate the answers: *GPT-3.5 Turbo* (Brown et al., 2020), *Phi-3-4k mini*[4] (3.82 B) (Abdin et al., 2024), and *Phi-3-4k-medium*[5] (14 B) (Abdin et al., 2024). We selected the Phi models as the Open LLMs, since, at the time of this study, they achieve the state-of-the-art performance in different tasks, such as language understanding, math, code, long context and logical reasoning (Abdin et al., 2024).[6] Due to the high computational resources, we limit the number of questions to 3000 per community in the SE-PQA dataset. Furthermore, due to budget constraints, for the closed-source GPT-3.5, we limit the number of questions to 1500. This results in approximately 100k answers for each prompt with the Phi-3-mini and Phi-3-medium, and around 50k

[4] microsoft/Phi-3-mini-4k-instruct

[5] microsoft/Phi-3-medium-4k-instruct

[6] Since none of these models details the source of their training data, we cannot exclude that the LLMs may have been trained also on the StackExchange data.

Table 1: Results on SE-PQA using data from Phi-mini.

Training	Model	P@1	NDCG@3	NDCG@10	MAP@100	λ
-	BM25	0.279	0.353	0.394	0.362	-
Real Answer	DistilBERT	0.253	0.329	0.378	0.344	-
	BM25 + DistilBERT	0.333*	0.415*	0.456*	0.421*	.3
Basic	DistilBERT	0.264*	0.340*	0.388*	0.352*	-
	BM25 + DistilBERT	0.336*	0.419*	0.460*	0.425*	.3
Personalized	DistilBERT	0.299*	0.374*	0.418*	0.384*	-
	BM25 + DistilBERT	0.345*	0.426*	0.465*	0.431*	.3
Contextual	DistilBERT	0.294*	0.369*	0.413*	0.379*	-
	BM25 + DistilBERT	0.342*	0.425*	0.464*	0.429*	.3

Table 2: Results on SE-PQA using data from Phi-medium.

Training	Model	P@1	NDCG@3	NDCG@10	MAP@100	λ
-	BM25	0.279	0.353	0.394	0.362	-
Real Answer	DistilBERT	0.253	0.329	0.378	0.344	-
	BM25 + DistilBERT	0.333*	0.415*	0.456*	0.421*	.3
Basic	DistilBERT	0.299*	0.373*	0.418*	0.383*	-
	BM25 + DistilBERT	0.344*	0.425*	0.465*	0.430*	.3
Personalized	DistilBERT	0.299*	0.374*	0.419*	0.384*	-
	BM25 + DistilBERT	0.347*	0.429*	0.468*	0.433*	.3
Contextual	DistilBERT	0.301*	0.375*	0.419*	0.385*	-
	BM25 + DistilBERT	0.338*	0.423*	0.463*	0.427*	.4

Table 3: Results on SE-PQA using data from GPT-3.5.

Training	Model	P@1	NDCG@3	NDCG@10	MAP@100	λ
-	BM25	0.279	0.353	0.394	0.362	-
Real Answer	DistilBERT	0.250	0.325	0.374	0.338	-
	BM25 + DistilBERT	0.328*	0.411*	0.451*	0.416*	.3
Basic	DistilBERT	0.256	0.327	0.372	0.340	-
	BM25 + DistilBERT	0.324*	0.404*	0.445*	0.410*	.4
Personalized	DistilBERT	0.264	0.335	0.382	0.348	-
	BM25 + DistilBERT	0.326*	0.407*	0.447*	0.413*	.4
Contextual	DistilBERT	0.264	0.335	0.382	0.349	-
	BM25 + DistilBERT	0.327*	0.408*	0.449*	0.414*	.4

answers with GPT-3.5.

2.2 Baselines

For the baseline, we follow the two-stage ranking architecture utilized by the SE-PQA dataset: the first stage is based on *BM25* (Robertson and Walker, 1994) with elasticsearch. We use the same BM25 hyperparameters of the SE-PQA paper. For the second stage we rely on *DistilBERT*[7] (Sanh et al., 2019). The document relevance score for the second stage is determined by a convex combination of the scores computed by BM25 and DistilBERT. The weights of the sum are optimized on the validation set by performing a grid search in the interval $[0, 1]$ with step 0.1. The optimal weights for each model are reported in the Section 3. In all the experiments, the second stage re-ranks the top-100 results retrieved by BM25. We fine-tune DistilBERT for 20 epochs, with a batch size of 128 and a learning rate of $5 \cdot 10^{-6}$ by using Triplet Margin Loss with in-batch random negatives, and a margin of $\gamma = 0.5$. We use AdamW as the optimizer and set the random seed to 42 for reproducibility purposes. We train and evaluate our models on a single A100 GPU.

3 Results

In this section, we present the outcomes of our experiments, evaluated using P@1, NDCG@3, NDCG@10 and MAP@100 as evaluation metrics. These metrics are crucial for assessing the precision and relevance of the retrieved documents. All metrics are computed using the ranx library (Bassani, 2022). Tables 1, 2 and 3 summarize the results for the models trained using data from Phi-mini, Phi-medium and GPT-3.5, respectively. In these tables, asterisks (*) indicate statistically significant improvements over the BM25 method, determined using a Bonferroni-corrected two-sided paired Student's t-test with 99% confidence. The λ column shows the optimized weight for BM25 during the second stage, with the weight for the neural model being $1 - \lambda$. We evaluate the models' effectiveness on the validation and test set from SE-PQA and thus their ability to retrieve human responses.

To compare performance between models trained on synthetic and real data, we train DistilBERT on human-written data, i.e., the real answers labelled as relevant for the user who wrote the question in the SE-PQA training set. Results of DistilBERT fine-tuned on human-written data are reported in the rows labelled *Real Answers*. From the three tables, we notice that neural reranker DistilBERT, trained on synthetic data, outperforms BM25 in a significant way, indicating that the synthetic dataset is suitable for training neural retrieval models. Indeed, DistilBERT improvement reaches 24% in P@1 and 18% in NDCG@10 over the BM25 model, when fine-tuned on the Personalized Answers. Even without adding contextual or personal information into the prompt, the model trained on *Basic* synthetic data achieves similar performances to the model trained on *Real Answers*. Furthermore, training the model on *Personalized* answers generated by both Phi mini and medium is able to improve remarkably over the model trained on the real answer. Regarding the data generated relying on GPT-3.5, it is worth noting that, even if the model trained on *Real Answers* achieves the best results compared to the model trained on synthetic data, the performance of all the models are in the same ballpark.

[7] distilbert/distilbert-base-uncased

Table 4: Detailed untruthful answers to the question reported in Section 3.1. *Model* and *Prompt type* refer to the LLM and prompt definition we use to generate the answers.

Question		In which book did Poirot find a four-foot long clue?
Model	Prompt type	Answer
	Real Answer	He found it in **The Murder on the Links**, and he made that statement in *Lord Edgware Dies*.
GPT 3.5	Basic	In Agatha Christie's book **"The Mystery of the Blue Train"**, Hercule Poirot finds a four-foot long clue.
	Personalized	The book in which Poirot found a four-foot long clue is **"The Labours of Hercules"** by Agatha Christie.
	Contextual	The book in which Poirot found a four-foot long clue is **"The Big Four"** by Agatha Christie. [...]
Phi-mini	Basic	The four-foot long clue you're referring to is found in Agatha Christie's novel **"The Mysterious Affair at Styles"**. [...]
	Personalized	In **"The A.B.C. Murders"** Hercule Poirot finds a four-foot long clue. [...]
	Contextual	The clue you're referring to is found in Agatha Christie's novel **"The Mysterious Affair at Styles"** However, it's important to note that the clue mentioned in the book is not exactly four feet long. [...]
Phi-medium	Basic	In the Agatha Christie novel **"The ABC Murders"** Hercule Poirot mentions a four-foot long clue. This clue is a piece of string that Poirot finds at the scene of one of the murders. [..]
	Personalized	The book in which Hercule Poirot found a four-foot long clue is **"The Mystery of the Blue Train"**.
	Contextual	In the Agatha Christie novel **"The ABC Murders"** Hercule Poirot finds a four-foot long clue. This clue is a piece of string that connects the crime scenes, which Poirot uses to unravel the mystery behind the serial killings. [...]

3.1 Exploratory Analysis

While LLMs have shown high capabilities in generating text, they are prone to hallucination, producing plausible yet incorrect statements (Tonmoy et al., 2024). This, combined with potentially outdated information, undermines the reliability of synthetic answers. Relying on such information increases the risk of spreading misinformation, particularly in communities where accurate information is critical, such as health, legal or finance. To assess the impact of hallucinations, we randomly select a total of 100 questions equally split among all communities, and manually check the synthetic answers.[8] We found that Phi-mini, Phi-medium, and GPT-3.5 generated incorrect answers in 41%, 35%, and 36% of cases, respectively.

In Table 4, we illustrate an example of hallucination in the Literature community. We ask the following question: "*In which book the famous fictional character Poirot finds a four-foot long clue?*" We note that, regardless of the prompt we generate, both Phi-3 and GPT-3.5 produce credible but incorrect answers. Although all the referenced books in the synthetic documents are novels in which Poirot is portrayed as the main character, none were a correct answer. Further examples of answers with hallucinations are provided in Appendix A.

Despite these challenges, as shown in Section 3, models trained on synthetic data can achieve better retrieval performance compared to the models trained on human-written answers.

[8]The subset of manually checked answers is available on the GitHub page.

4 Conclusion and Future Works

In this study, we explore the generation of synthetic data for personalized information retrieval (PIR) tasks using Large Language Models (LLMs). By leveraging the SE-PQA dataset, we generate synthetic answers and propose a new dataset, Sy-SE-PQA, with varying levels of personalization and contextual information, employing models like GPT-3.5 Turbo, Phi-3-mini, and Phi-3-medium. Our experiments show that neural re-rankers, in particular, DistillBERT, fine-tuned on synthetic data, significantly outperformed the traditional BM25 method. This indicates that synthetic datasets can effectively train neural retrieval models, enhancing their ability to provide relevant and precise answers. Our findings highlight the potential of synthetic data in improving PIR tasks, especially when personalized and context-aware responses are crucial. The notable improvements in metrics over the baselines underscore the value of incorporating personalized information into the answer generation process. Furthermore, the exploratory analysis revealed challenges related to hallucinations in generated answers, emphasizing the need for continued refinement of LLMs to ensure accuracy and reliability. We leave as future works the development of more complex prompt techniques that could exploit a wider range of user-related and contextual features available in the SE-PQA dataset, which are not used in our current models. By doing so, we hope to enhance the accuracy and relevance of generated answers, further improving the robustness of our retrieval models.

5 Limitations

Despite the promising results, our study has several limitations that need to be discussed:

Bias and Fairness Furthermore, an ethical concern associated with the content generated with LLMs is bias and fairness. The LLMs are trained on large datasets that may contain biases reflecting societal prejudices. These biases can be, inadvertently, perpetuated by utilizing data generated by these LLMs, which could lead to unfair or discriminatory outcomes.

Computational and Budget Constraints The generation of synthetic data using LLMs is resource intensive. Due to the computational and budget constraints, we limited the number of questions we could generate the answers to, affecting potentially the diversity of the generated dataset.

Evaluation Metrics In this work we relied on the classical IR metrics such as P@1, NDCG@3, NDCG@10 and MAP@100, which certainly are useful to evaluate the retrieval effectiveness, but they may not capture completely the quality of the generated answers, for example, readability, user satisfaction and coherence.

Generalization to other domains We tested the data generation procedure in the specific task of question answering. The effectiveness of the proposed methods needs to be tested also in other domains to explore their adaptability.

While the synthetic data holds great potential, addressing these limitations is very important for advancing the field, while ensuring the robustness and applicability of the models in real-world scenarios.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. A test collection of synthetic documents for training rankers: Chatgpt vs. human experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5311–5315, New York, NY, USA. Association for Computing Machinery.

Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for aol searcher no. 4417749. *New York Times*, 9(2008):8.

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *European Conference on Information Retrieval*, pages 259–264. Springer.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A context-aware time model for web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 205–214.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, et al. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. Se-pqa: Personalized community question answering. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1095–1098.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, pages 1–es.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mirco Speretta and Susan Gauch. 2005. Personalized search based on user search histories. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 622–628. IEEE.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Chen Wu, Ming Yan, and Luo Si. 2017. Ensemble methods for personalized e-commerce search challenge at cikm cup 2016. *arXiv preprint arXiv:1708.04479*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

A Appendix A

Table 5: Example of hallucinations in the History community

Question Title	Which figure is this a statue of?	
Question Body	In Ocean's 12, [...] we see this statue in the foreground. My question is, what is this statue or rather or who is this a statue of? [...]	
Model	Prompt type	Answer
Human written		It is probably a statue of **Antinous. (Or Hermes)**. In fact it is probably this one here (notice the object behind his leg). It is located in the Capitoline musuem in Rome. It is supposed to be Antinous in the shape of the Greek god Hermes.
GPT 3.5	Body pers.	The statue in question is a depiction of **Laocoon and His Sons**. It is a marble sculpture from the Hellenistic period of ancient Greece and is now on display at the Vatican Museums in Rome. [...]
	Body base	The statue in Ocean's 12 that you are referring to is of the Greek goddess **Artemis**, also known as Diana in Roman mythology. [...]
	Body context	The statue seen in the movie Ocean's 12 is a depiction of the ancient Greek mythological figure, **Laocoon**. He was a Trojan priest who warned his people against accepting the wooden horse from the Greeks during the Trojan War. [...]
Phi-mini	Body pers.	The statue you are referring to in the Ocean's 12 scene is likely a representation of the Greek goddess **Athena**. [...]
	Body base	The statue you're referring to in the Ocean's 12 scene is likely a representation of the Greek goddess **Athena**. [...]
	Body context	The statue you're referring to in "Ocean's 12" is likely a representation of the Greek goddess **Athena**. [...]
Phi-medium	Body pers.	The statue in the movie Ocean's 12 is a representation of the Greek mythological figure **Medusa**. [...]
	Body base	The statue you are referring to in the movie Ocean's 12 is a representation of the Greek goddess **Aphrodite**, also known as Venus in Roman mythology. [...]
	Body context	The statue you are referring to in the movie Ocean's 12 is a representation of the Greek goddess **Athena**.[...]