An Explainable LLM Approach for News Articles

Gabriel Hili[1], Dylan Seychell[1], and Konstantinos Makantasis[1]

Department of Artificial Intelligence, University of Malta, Malta
{gabriel.hili,dylan.seychell,konstantinos.makantasis}@um.edu.mt

Abstract. News consumption plays an important role in our daily lives. This paper introduces Malta-7B, an LLM-driven semantic search engine and comprehension tool for news article retrieval, which includes explainability by design. Explainability is built into the system by design, and through a detailed user study with 11 participants, we find that Malta-7B sufficiently communicates the model's rationale to the user via the provided source listings. The system maintains two LLMs, one for comprehension and one for news article retrieval. These two modules work in tandem to answer the user's queries and provide relevant sources. Additionally, Malta-7B can also be run on consumer-grade hardware using open-source LLMs, allowing the system to be publicly accessible without the need for API subscriptions. We open-source Malta-7B under the MIT license on GitHub.

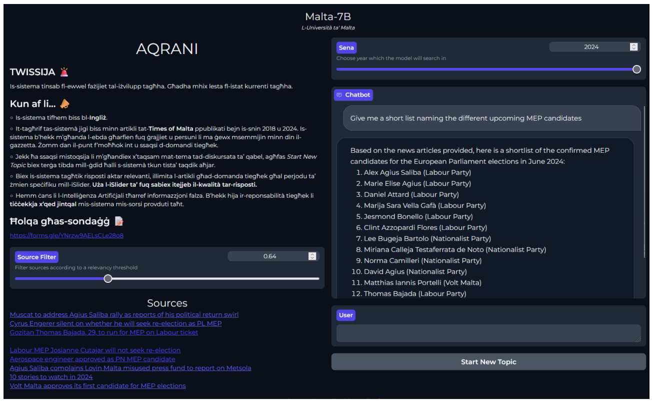
Keywords: Explainability · News Analysis · Natural Language Processing · Large Language Models · Retrieval-Augmented Generation · Media Studies.

1 Introduction

News consumption is integral to receiving information about the world around us. However, a study by the European Union indicated that 37% of all Europeans in 2023 used social media as a source of their news consumption [13]. This number rises to a staggering 70% for the Maltese population. Unfortunately, social media platforms are susceptible to echo chambers and feedback-reward loops, which makes them inappropriate for ethical news consumption. For those who access online news portals directly, while many include a search feature to improve retrieval, these suffer from the same disadvantages as traditional lexical search, such as the inability to handle misspellings, synonyms, or polysemies (words that have a different meaning depending on the context).

As a result, our work addresses these setbacks by introducing Malta-7B: an LLM-driven semantic search engine and comprehension tool for news article retrieval, with a foundation rooted in explainability. Such a system considers contextualised information in the user query for retrieving relevant news articles, unlike searches commonly found in online news portals. Moreover, unlike content recommendation algorithms on social media platforms, user profiling does not affect news article retrieval, eliminating the possibility of an echo chamber. Malta-7B is a first of its kind in the media analysis landscape and can potentially introduce a new paradigm of news consumption, as indicated in this user study.

Since Large Language Model (LLM) hallucinations threaten to deceive the reader with misinformation, these must be nipped in the bud. Explainability is an integral part of the system and is evaluated in Section 4.1. In fact, 81.8% of participants (9/11) stated that the system's overall design was explainable when evaluating a predetermined news topic of their choice, with 55.6% of those (5/9) giving it the highest score possible. Local explainability is built into the system by design by grounding the LLM outputs on ground-truth user-verifiable sources, wholly bridging the LLM black box gap between the newspaper and the end-user. All relevant sources are provided alongside every LLM response to promote explainability, interpretability, and ethical news consumption.



TWISSIJA 🚨

Is-sistema tinsab fl-ewwel fażijiet tal-iżvilupp tagħha. Għadha mhix lesta fl-istat kurrenti tagħha.

Kun af li... 📣

- Is-sistema tifhem biss bl-**Ingliż**.
- It-tagħrif tas-sistemà jigi biss minn artikli tat-**Times of Malta** ppublikati bejn is-snin 2018 u 2024. Is-sistema b'hekk m'għanda l-ebda għarfien fuq ġrajjiet u persuni li ma ġewx msemmijin minn din il-gazzetta. Żomm dan il-punt f'moħħok int u ssaqsi d-domandi tiegħek.
- Jekk ħa ssaqsi mistoqsija li m'għandiex x'taqsam mat-tema tad-diskursata ta' qabel, agħfas *Start New Topic* biex terġa tibda mill-ġdid ħalli s-sistemà tkun tista' taqdik aħjar.
- Biex is-sistema tagħtik risposti aktar relevanti, illimita l-artikli għad-domanda tiegħek għal perjodu ta' żmien speċifiku mill-iSlider. **Uża l-iSlider ta' fuq sabiex itejjeb il-kwalità tar-risposti.**
- Hemm ċans li l-Intelliġenza Artifiċjali tħarref informazzjoni falza. B'hekk hija ir-reponsabilità tiegħek li **tiċċekkja x'qed jintqal** mis-sistema mis-sorsi provduti taħt.

Ħolqa għas-sondaġġ 📝

https://forms.gle/YNrzw9AELsCLe28o8

Filter sources according to a relevancy threshold

Sources

Muscat to address Agius Saliba rally as reports of his political return swirl
Cyrus Engerer silent on whether he will seek re-election as PL MEP
Gozitan Thomas Bajada, 29, to run for MEP on Labour ticket

Labour MEP Josianne Cutajar will not seek re-election
Aerospace engineer approved as PN MEP candidate
Agius Saliba complains Lovin Malta misused press fund to report on Metsola
10 stories to watch in 2024
Volt Malta approves its first candidate for MEP elections

Choose year which the model will search in

Give me a short list naming the different upcoming MEP candidates

Based on the news articles provided, here is a shortlist of the confirmed MEP candidates for the European Parliament elections in June 2024:

1. Alex Agius Saliba (Labour Party)
2. Marie Elise Agius (Labour Party)
3. Daniel Attard (Labour Party)
4. Marija Sara Vella Gafà (Labour Party)
5. Jesmond Bonello (Labour Party)
6. Clint Azzopardi Flores (Labour Party)
7. Lee Bugeja Bartolo (Nationalist Party)
8. Miriana Calleja Testaferrata de Noto (Nationalist Party)
9. Norma Camilleri (Nationalist Party)
10. David Agius (Nationalist Party)
11. Matthias Iannis Portelli (Volt Malta)
12. Thomas Bajada (Labour Party)

Fig. 1. A sample of Malta-7B User-Interface

Towards that direction, at the core of the proposed system, there are two language models: one for retrieval (R-Model) and one for comprehension (C-Model). Specifically, we utilise Nomic AI's `nomic-embed-text-v1` [12] and Mistral's `Mistral-7B-Instruct-v0.2` [7] with GPTQ quantization [3] respectively. A regularly updated vector database contains news articles embedded by the R-Model, which are promptly retrieved using Hierarchical Navigable Small World (HNSW) [10] indexing based on the embedded user queries. Top-k retrieved news articles are filtered by a cosine distance threshold to the user query (s) and passed as non-parametric knowledge to the C-Model, further filtering irrelevant text per the global system prompt and the user's instructions. Both language models' behaviours are explained via a source list at the end of each iteration. Sources are determined by calculating the cosine similarity between the C-Model's output and each retrieved article. Figure 2 shows a comprehensive system diagram.

This project is a multi-disciplinary work that combines LLMs, eXplainable AI (XAI), and media analysis to provide a see-through news article search and comprehension tool, with evidence from our user study supporting a 27% preferability over traditional manual search for widely discussed news topics. The rest of this paper is structured as follows: Section 2 covers related work, Section 3 discusses the implementation details of Malta-7B, and Section 4 evaluates the user study and outlines the system's limitations.

2 Related Work

2.1 News Consumption

News consumption has evolved hand-in-hand with advancements in technology, from Gutenberg's printing press all the way to the Internet. With the emergence of LLMs, a new paradigm of news consumption is opening up. For example, NewsGPT.ai [1] is a 24/7 LLM-powered news channel that generates summaries of various news events and also incorporates image generation and AI avatars. However, a drawback to this system is that it is only limited to international mainstream news, making it hard to follow slowly unfolding events from a particular country. Moreover, no ground-truth verifiable sources are included with the LLM-generated summaries, relying instead on crowdsourcing techniques, which are prone to tyranny of the majority. Another LLM-based news system, similarly named NewsGPT-Clickbait-Buster [2] utilises `gpt-3.5-turbo` to search and comprehend news articles similar to Malta-7B. However, the system is limited to searching news published within the last 30 days, which restricts the analysis of older news content. Moreover, while the above-mentioned news systems utilise LLMs effectively, none address the risk of LLM hallucinations on unsuspecting news readers, nor do they include an explainability report evaluating the different aspects of the system.

2.2 Measuring Explainability in LLMs

LLMs have become the state-of-the-art technique for a wide range of application due to their emergent and zero-shot learning capabilities. However, the main drawback of LLMs is that it is difficult to determine the sequence of steps taken to arrive at a conclusion (interpretability) as well as demystifying the rationale behind the outputs especially to a non-technical audience (explainability). This problem is further amplified in cases where LLMs scale to hundreds of billions of parameters. Hence, common techniques such as gradient-based methods [16] or

[1] https://newsgpt.ai/

[2] https://github.com/timho102003/NewsGPT

SHAP values [9] do not scale well with the increasing parameter count of LLMs.

A core part of Transformer [17] based models including LLMs is attention [1]. By visualizing the attention between input and output tokens, insights into explaining the model's outputs can be deduced. However, there has been much debate regarding the usefulness of attention for accurately gauging model interpretability and explainability. In-fact, many have argued that attention is not a descriptive enough representation of neither interpretability or explainability [15][6]. However, since LLMs are mostly sought for their strong emergent properties, explainability techniques rarely ever target their complex inner-workings but rather aim to derive conclusions by exhaustively prompting the model over varying system-prompts or fine-tuning datasets [18]. Due to the context sensitive nature of news consumption as well as the long detailed format of news articles, we opt out of implementing a visualisation-based tool for measuring explainability on the token-space.

LLM confabulations, which are also popularly known as hallucinations, refer to unjustified responses or beliefs which contain false information but are presented as plausible facts [2]. The lack of relevant data and duplicate data in the training dataset are the main offenders in hallucinating LLMs [18]. In-fact it has been shown that having 10% of the training data be redundant/duplicate can degrade the performance of an 800M parameter language model to that of 400M parameters [5]. We tackle similar experiences during the development of our system which we further elaborate on in Section 4.

One of the research challenges currently faced by LLM explainability techniques is that ground-truth explanations are usually inaccessible to the end-researcher, which makes testing the effectiveness of such algorithms difficult [18]. Moreover, the lack of ground-truth makes measuring faithfulness of the LLM responses to the underlying sources problematic. Since our system primarily makes use of the retrieved context to answer the fact-sensitive news article queries, parametric knowledge is rarely ever invoked in the comprehension model, circumventing the necessity for ground-truth explanations for this aspect of the system. With respect to non-parametric knowledge, sources are provided by design as the system is grounded on publicly available real-world newspaper articles which the end-user can use to verify important facts. The design of the system also makes measuring the lexical and semantic faithfulness of the comprehension model output to the ground-truth sources trivial.

In conclusion, while the field of LLM explainability is broad and multifaceted, there exist many signs that indicate the field is still in its infancy. Explainability techniques vary according to the ultimate downstream use case of the base model and often necessitate clever exploitation of all available components of the system. Moreover, since explainability techniques concern themselves with the understanding of the underlying system, especially to non-technical audiences, surveys remain few of the best ways to evaluate such techniques.

3 Implementation

Malta-7B is a news search and comprehension engine composed of three main components, the Retrieval Model (R-Model), the Comprehension Model (C-Model), and the vector database (See Fig. 2). The C-Model's function is to comprehend the users' instructions and present them with the requested information from the news articles provided by the R-Model. The R-Model's function is to embed user queries, news articles, and the C-Model's responses in order to facilitate Retrieval Augmented Generation (RAG) [8] between the C-Model and the vector database, as well as generate the relevant source listings for each of the C-Model's response. All relevant code for Malta-7B is open-sourced on GitHub [3] under the MIT license.

Malta-7B's UI is built using `gradio==3.36.1` and is composed of the following components (See Fig. 1):

Component	**Description**
Disclaimer and usage tips	Currently only written in Maltese.
Chatbot interface	Includes chat history and a prompt input field.
New Topic Button	Clears the current chat history and sources.
Year selection slider	Filters article retrieval by chosen year
Source filter slider	Filters sources by their cosine distance to responses.
Source list	Lists all relevant sources.

Table 1. Malta-7B individual UI components

3.1 Populating Vector Database

The initial preliminary step before hosting the system is the population of the vector database. News articles, as well as useful meta-data, are formatted into a single text block and embedded wholly by the retrieval model. We utilise the open-source `nomic-embed-text-v1` [12] text embedder for its 8192-token sequence length as well as its impressive performance in MTEB [11] (62.39), LoCo [14] (85.53) and Jina Long Context [4] (54.16). The R-Model's long sequence length is crucial for news article embedding, as this circumvents the need for document chunking, removing important descriptive keywords crucial to accurate topic clustering. We noticed that inserting a system prompt for the R-Model before the text improves ultimate downstream performance during retrieval (See

[3] https://github.com/GabrielFreeze/Malta-7B

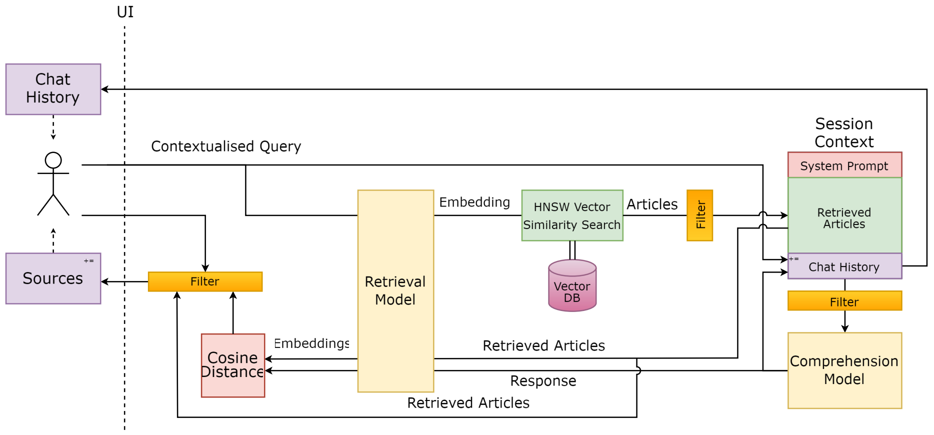


Fig. 2. System Diagram

R-Model System Prompt
Represent this news article for searching relevant passages about events, people, dates, and facts. Make explicit reference to the date, author, and title of the news article. Prioritize recent data.

C-Model System Prompt
You are a helpful assistant with knowledge of published Maltese news articles. Your main task is to answer Malta-related news queries while/or maintaining coherent dialouge with the user. You will not make up information. You will not reference articles unrelated to the query. If the query does not match any of the articles provided, you will not attempt to answer. Do NOT include URLs in your answer. You will keep your answers short and concise. You can do this, I believe in you.

Table 2. System prompts for the R-Model and C-Model respectively.

Table 2).

Using this model, we insert 44,230 publicly available Times of Malta [4]articles spanning over six years (01/01/2018 - 13/03/2024) into the vector database. Articles that exceed 2048 tokens using the C-Model's tokenizer are discarded so as not to overwhelm Malta-7B's context length with a single article.

3.2 Relevant Article Retrieval

The concatenation of the user's current chat messages is passed to the R-Model to be used as the search query for the vector database. Due to the time-sensitive

[4] http://www.timesofmalta.com

nature of news articles, we limit news article searches to only a year's worth of data as specified by the user through the slider. Although this limits the system from analysing news stories that slowly unfold throughout the years, in general this improves retrieval performance as it lessens the vast number of irrelevant articles that would have to be sieved through by HNSW [10] vector search. The top-k relevant news articles to the query are retrieved from the vector database, and all articles with a cosine similarity score to the user query less than some threshold s are discarded. This threshold ensures that if no relevant news articles exist for the provided query, none will be passed as context to the C-Model to mitigate the possibility of hallucinations. During our user study, we set $k = 35$ and $s = 0.55$.

3.3 Presenting Relevant Information

The filtered relevant articles are inserted between the C-Model's system prompt and the current chat history. Through our testing, we found it best to maintain all relevant articles as a single aggregate at the beginning of the input to the C-Model, rather than appending them with the corresponding user query that prompted their retrieval. This makes it easier for the C-Model to continue with the flow of the conversation since the user-assistant chat dialogue is not delimited with large quantities of news article excerpts.

We utilise `TheBloke/Mistral-7B-Instruct-v0.2-GPTQ` for the C-Model due to its better performance over larger LLMs such as LLaMA-2-13B-chat. Although larger models such as `Mixtral-8x7B-Instruct-v0.1` could be run with a tolerable token generation speed given the provided hardware (RTX4090 24GB), due to the large number of context-relevant articles provided to the C-Model, the generation speed would become unacceptable. Reducing k to limit the number of retrieved articles to reduce the length of the input sequence would then risk omitting important news articles related to the user's query. Moreover, since many participants would be familiar with the token generation speed of popular cloud-based chatbot services, we preferred the quantized Mistral-7B's fast response times.

Given the provided relevant news articles, the C-Model considers the system prompt and previous chat history to fulfil the user's query and maintain coherent dialogue. The system prompt instructs the model to follow a constitution that discourages it from responding with information not found in the provided news articles (See Table 2). We use `exllama` to increase the context length of Mistral-7B to 32,768 tokens as the original 4096 token sequence length greatly limits the amount of news articles Malta-7B can process at any given time. Output tokens are streamed back to the user, and once the EOS (End Of Sentence) token is generated, the source list is generated. From the sources provided by the R-Model, the C-Model may not have consulted all news articles as they may have been irrelevant to the specific instructions of the user query. Moreover, prompting the model to specify which news articles were consulted led to suboptimal results more often than not. Hence, to filter out the unused articles from the C-Model's output, we compute the cosine similarity between the C-Model's response and each of the R-Model's provided news articles. The intuition is that heavily consulted sources will result in the highest cosine similarity. Users can filter through these sources by varying the threshold via a provided slider.

3.4 Summary

Malta-7B is an open-source news search and comprehension engine composed of a Retrieval Model (R-Model), a Comprehension Model (C-Model), and a vector database. The R-Model embeds user queries, news articles, and the C-Model's responses for RAG [8] and generates relevant source listings. The C-Model follows user instructions and presents requested information from the relevant news articles provided. The system's UI is built with `gradio==3.36.1` and features a chatbot interface, a year selection slider, and a source filter slider. Around 45,000 Times of Malta articles are embedded using the R-Model and inserted into the vector database. The user filters their search query by year, from which relevant news articles are retrieved using HNSW [10] vector search. Articles not referenced by the C-Model are filtered based on cosine similarity between the C-Model's responses and the R-Model's provided sources, allowing users to access relevant sources more efficiently. Lastly, Malta-7B is optimized for consumer-grade hardware (RTX 4090 24GB), resulting in negligible waiting times and satisfactory performance.

4 Evaluation

	Question	1^{st} Topic						2^{nd} Topic					
		1	2	3	4	5	6	1	2	3	4	5	6
1	How relevant were Malta-7B's responses to your query?	-	-	-	2	3	6	-	-	2	1	5	3
2	How factual were Malta-7B's responses?	-	-	1	2	4	4	-	2	1	-	5	3
3	Where the provided sources useful to your topic of discussion?	-	-	-	1	1	9	-	1	-	3	2	5
4	Did the retrieved sources serve as a lens into the model's rationale?	-	-	-	1	6	4	-	1	1	2	3	4
5	Did you find the system overall to be explainable?	-	-	-	2	4	5	-	1	1	2	2	5
		Yes		Both		No		Yes		Both		No	
6	Would you use this system over traditional manual search?	3		8		-		2		7		2	

Table 3. Overview of the results from the user-study. Numbers indicate total participants that chose the corresponding option.

4.1 Limitations

The main limitations and drawbacks of the system are as follows.

Restricted Search Malta-7B can only handle a year's worth of data at any given time. Since many past events are usually of little relevance to the user, having an unrestricted search over all possible news articles risks presenting outdated knowledge to the user. Queries such as "*electric vehicle grants*" or "*who is the current prime minister*" are prone to this issue. Moreover, while the C-Model can filter out outdated news articles (provided it has a sufficient parameter count), this still clogs the C-Model's context length, which could otherwise be used for more relevant articles pertaining to the specific year requested. We see two possible ways to mitigate this limitation. Either utilise a C-Model with a better capability at distinguishing events chronologically or implement another filter using a language model to distinguish between events based on the time frame requested by the user.

Single Newspaper Currently, Malta-7B works best when the articles in the vector database are all from a single source, in this case, the Times of Malta. We found that including more than one newspaper significantly degrades performance during testing. While different newspapers ensure that a wide range of opinions and perspectives are at the R-Model's disposal, many newspapers cover the same events with no additional explicit information. As a result, this clutters the C-Model's context length with duplicate information, significantly degrading its performance. A way to counteract such a limitation is to deduplicate the retrieved articles, keeping in mind that the document pool should not be biased towards a specific newspaper. Otherwise, another search tuning parameter can be added to the UI, indicating which newspapers the user would like to search for.

4.2 User Study

To evaluate the explainability of Malta-7B, as well as general performance, we conducted a focused user study on 11 participants (7 males, 4 females) where they made use of the system twice and gave corresponding feedback (See Table 3). Ten (10) participants were senior (4yrs+) University of Malta Students from a variety of disciplines, namely Law (3), Architecture (2), and Artificial Intelligence (5), whilst the remaining participant (1) was studying nursing. The questionnaire prompts participants to use the system for as long as they like, then return to the survey once they confirm they have used it enough to give valuable feedback. Many of the questions are answered with a linear scale with values ranging from 1 (lowest) to 6 (highest).

The user study prompted participants to make use of the system twice: first on a predetermined topic of their choosing and secondly, on any news topic they wish (See Tables 4 and 5). The reason for prompting users to pick from a set of predetermined topics is that Malta-7B works best on topics with substantial coverage in the media. For topics not extensively covered in news articles, queries requesting certain information not explicitly found in any article can lead to hallucinations and incorrect sources. This can be trivially fixed by including a more comprehensive range of news articles, however for the scope of our study we wanted to collect usage data on topics the system is better suited for. Regardless, participants were still free to form their query however they wished as long as it belonged to the overall predetermined topic.

Topic	**(%)**
Car Crashes in Locality	36.4% (4)
Corradino Building Collapse	27.3% (3)
Daphne Caruana Galizia	18.2% (2)
Upcoming MEP Elections	9.1% (1)
Information about Protests	9.1% (1)

Table 4. Predetermined topics picked by participants to test Malta-7B on.

Topics
“*Journalisim in Malta*”, “*Government Price Decreases*”, “*Musical events*”, “*Palestine comments by Maltese politicans*”, “*The court cases of Bryan Tonna and Nexia BT* ”, “*cats for adoption*” “*Smoking in Malta*”, “*The Malta Security Service*”,“*Pre-Insolvency Act*”,“*Gozo*”,“*Maltese sport*”

Table 5. Topics chosen by participants to test Malta-7B on.

From this user study, the majority of participants found Malta-7B's responses to be relevant and factual (a score of 5 or higher) for both their tries with the system (Q.1, Q.2). Moreover 9 out of 11 participants found the provided sources to be very useful to the predetermined topic of their choice, giving the highest score possible (Q.3). Participants were more inclined to give a lower score for this question in their second run with a topic of their choice. However, the results still indicate that the retrieved sources were relevant.

The next part of the survey aims to evaluate the explainability of the system. The following short definition of explainability was provided before Question 4: “*Explainability means understanding the rationale of the system's output from a non-technical perspective.*”. The majority of participants gave a score of 5 or higher when asked if the retrieved sources managed to serve as a lens into the model's rationale (Q.4). Around 91% of participants gave such a score for the first topic, however this number dropped to 64% when evaluating for the second topic. When asked if they found the overall system to be explainable after being provided with the above definition, 82% gave a score of 5 or higher for the first topic and 64% for the second topic (Q.5). Lastly, all participants stated they would use the system when researching about there first topic, with around 27% preferring to use the system over traditional manual search. For the second topic, 9 out of 11 participants stated they would use the system during their research.

4.3 Post-Study Feedback

After using the system twice and answering the relevant questions, we ask participants to provide their thoughts on the overall system. 8 out of 11 participants gave a score of 5 or higher when asked to rate the performance of Malta-7B in answering their queries. On the other hand, around 37% of participants (4) gave a score of 5 or higher when asked whether it was easy to spot deception/hallucinations, with another 37% giving a score of 4. Although, in theory, all relevant articles are provided to the user so they can fact-check important evidence, this is not immediately apparent to the user and requires them to sieve through multiple sources. Lastly, 10 out of the 11 participants stated they see themselves using an improved system version with image and topic integration as part of their primary research procedure, with the remaining stating *maybe.*

4.4 Main findings

The main findings of the user study are as follows. Most participants found Malta-7B's responses relevant and factual for both predetermined topics and topics of their choice. Moreover, the majority of participants found the sources to be useful to their search query. With regards to explainability, participants found the system to exhibit promising signs of explainability, indicating that Malta-7B's design is sufficient for effectively communicating its decision-making process to the users, even to those without technical expertise. Lastly, almost all participants indicated they would be willing to integrate malta-7B into their news research procedure, with 27% preferring it over the traditional manual search for their pre-determined topic of choice.

5 Conclusion

In this work, we introduced Malta-7B, an explainable LLM-driven news search and comprehension engine. Our system employs a comprehension model and a retrieval model in order to retrieve relevant news articles and present the information according to the user's instructions. Sources are provided with every response so that the user can verify any important information. Malta-7B also features a slider to filter the articles to be retrieved by year and a slider to filter only the most relevant sources. We evaluate the performance and explainability of our system via a user study (11 participants) and find that the majority find Malta-7B to be an explainable useful tool they wish to integrate into their news research/consumption methodology. Future work will focus on scaling the user evaluation further to achieve statistically significant results from the focused, in-depth study presented in this paper.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity (2023)
3. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)
4. Günther, M., Ong, J., Mohr, I., Abdessalem, A., Abel, T., Akram, M.K., Guzman, S., Mastrapas, G., Sturua, S., Wang, B., Werk, M., Wang, N., Xiao, H.: Jina embeddings 2: 8192-token general-purpose text embeddings for long documents (2024)
5. Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C., Amodei, D., Joseph, N., Kaplan, J., McCandlish, S.: Scaling laws and interpretability of learning from repeated data (2022)
6. Jain, S., Wallace, B.C.: Attention is not explanation. arXiv preprint arXiv:1902.10186 (2019)
7. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021)
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
10. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs (2018)
11. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark (2023)
12. Nussbaum, Z., Morris, J.X., Duderstadt, B., Mulyar, A.: Nomic embed: Training a reproducible long context text embedder (2024)
13. Parliament, E., for Communication, D.G.: Media news survey 2023. European Parliament (2023). https://doi.org/doi/10.2861/11595
14. Saad-Falcon, J., Fu, D., Arora., S.: Long-context retrieval models with monarch mixer (2024)
15. Serrano, S., Smith, N.A.: Is attention interpretable? arXiv preprint arXiv:1906.03731 (2019)
16. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey (2023)