applied sciences

MDPI

Article

Analysis of the Effectiveness of Model, Data, and User-Centric Approaches for Chat Application: A Case Study of BlenderBot 2.0

Chanjun Park [1,†], Jungseob Lee [2,†], Suhyune Son [2,†], Kinam Park [3], Jungsun Jang [4,*] and Heuiseok Lim [2,*]

[1] Upstage, Yongin-si 16942, Republic of Korea; chanjun.park@upstage.ai
[2] Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea; omanma1928@korea.ac.kr (J.L.); ssh5131@korea.ac.kr (S.S.)
[3] Human-Inspired AI Research, Korea University, Seoul 02841, Republic of Korea; spknn@korea.ac.kr
[4] Department of Korean History, Korea University, Seoul 02841, Republic of Korea
* Correspondence: empyrean@korea.ac.kr (J.J.); limhseok@korea.ac.kr (H.L.)
† These authors contributed equally to this work.

Abstract: BlenderBot 2.0 represents a significant advancement in open-domain chatbots by incorporating real-time information and retaining user information across multiple sessions through an internet search module. Despite its innovations, there are still areas for improvement. This paper examines BlenderBot 2.0's limitations and errors from three perspectives: model, data, and user interaction. From the data perspective, we highlight the challenges associated with the crowdsourcing process, including unclear guidelines for workers, insufficient measures for filtering hate speech, and the lack of a robust process for verifying the accuracy of internet-sourced information. From the user perspective, we identify nine types of limitations and conduct a thorough investigation into their causes. For each perspective, we propose practical methods for improvement and discuss potential directions for future research. Additionally, we extend our analysis to include perspectives in the era of large language models (LLMs), further broadening our understanding of the challenges and opportunities present in current AI technologies. This multifaceted analysis not only sheds light on BlenderBot 2.0's current limitations but also charts a path forward for the development of more sophisticated and reliable open-domain chatbots within the broader context of LLM advancements.

Keywords: BlendorBot; chatbot; dialogue; deep learning; natural language processing

check for updates

Citation: Park, C.; Lee, J.; Son, S.; Park, K.; Jang, J.; Lim, H. Analysis of the Effectiveness of Model, Data, and User-Centric Approaches for Chat Application: A Case Study of BlenderBot 2.0. *Appl. Sci.* **2024**, *14*, 4821. https://doi.org/10.3390/app14114821

Academic Editor: João M. F. Rodrigues

Received: 20 April 2024
Revised: 26 May 2024
Accepted: 31 May 2024
Published: 2 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Developing agents capable of conversing on any topic represents a significant challenge in the field of open-domain dialogue systems. Such capability is a crucial aspect of human intelligence in natural language understanding and finds applications across a wide array of industrial services. The ultimate objective of this research is to create AI that can engage in meaningful dialogue, providing responses that are both relevant and engaging to any user.

Recent studies, including [1,2], have proposed dialogue modeling approaches aimed at generating knowledge-based responses. Additionally, research such as [3,4] has introduced dialogue modeling techniques for generating empathetic responses, tailored to the user's interests. These advancements have notably enhanced the performance of conversational models, enabling them to generate responses by leveraging knowledge and understanding users' interests. However, achieving the desired level of conversational skill remains a challenge.

One major limitation is the difficulty of incorporating recent information into the dialogue model, which typically relies on static knowledge from datasets collected at a specific point in time. To accurately reflect the dynamic nature of real-world knowledge, conversational models would need to continually update their datasets, a process that is both resource-intensive and challenging. A model that can adapt to changing knowledge in real-time would more closely mimic human-like conversation, deepening engagements with users and advancing the goal of human–AI interaction.

Another issue is the short-term memory of dialogue models compared to human memory. While humans can remember details about others for extended periods, current models tend to forget user-specific information, such as hobbies, after a limited number of turns in a conversation. This results in the need for users to repeat information, reducing the diversity and richness of conversations and potentially leading to user disengagement.

BlenderBot 2.0, as discussed in [5,6], addresses these challenges by incorporating multi-session memory and internet search capabilities. This approach allows the model to maintain interest and summarize conversations over longer periods, achieving a semblance of long-term memory. Moreover, it can update its knowledge base in real time through internet searches. However, BlenderBot 2.0 still faces significant challenges, including reliance on inaccurate information from the internet and generating incorrect search queries.

This paper examines the limitations of BlenderBot 2.0 and, by extension, current open-domain dialogue models from three perspectives: model, data, and user. It proposes improvements beyond conventional modeling techniques, aiming to address errors and enhance the dialogue model's performance across various metrics.

The significance of this research lies not only in advancing the technological capabilities of open-domain dialogue systems but also in its broader implications for human-AI interaction. By addressing the limitations of current models, we move closer to creating AI that can understand and adapt to the dynamic nature of human knowledge and memory. This advancement has the potential to revolutionize the way we interact with machines, making AI more accessible, personalized, and effective in a variety of applications, from education and healthcare to customer service and beyond. Ultimately, this research contributes to the development of AI that can more deeply and meaningfully engage with humans, fostering a future where AI supports and enhances our daily lives in increasingly sophisticated ways.

This article is organized as follows: the introduction provides an overview of the research objectives and the significance of developing advanced open-domain chatbots. Following this, the section on related works and background explores the evolution of chatbot technologies, with a particular focus on BlenderBot 1.0 and 2.0. The subsequent sections present a detailed error analysis from three perspectives: model-centric, data-centric, and user-centric. For each perspective, we identify key limitations and propose potential improvements. The discussion section synthesizes these findings and offers comprehensive strategies for enhancing the performance of BlenderBot 2.0 and similar systems. Finally, the conclusion summarizes the main contributions of this study and outlines directions for future research.

2. Related Works and Background

2.1. Open-Domain Chatbot

Chatbots are systems engineered for engaging in extended conversations, designed to replicate the unstructured nature of chats characteristic of human-to-human interaction [7]. Essentially, a chatbot system produces a response R, taking into account the user's message M and the history of the dialogue C. The primary objective of such systems is to generate responses that are indistinguishable from those a human might provide during a conversation [8]. Unlike task-oriented dialogue systems, which are limited to discussing a predetermined set of topics, developing an open-domain chatbot poses a greater challenge due to the necessity of generating responses across a broad spectrum of everyday topics.

The evolution of chatbot services began with the pioneering chatbot Eliza [9] and has since expanded to include modern examples such as Apple's Siri (https://www.apple.com/kr/siri/ (accessed on 1 May 2024)), Microsoft's XiaoIce (also known as Little Bing) [10] (https://www.xiaoice.com/ (accessed on 1 May 2024)), Simsimi (https://www.simsimi.com/ (accessed on 1 May 2024)), and Lee-Luda (https://luda.ai/ (accessed on 1 May 2024)), showcasing the global and diverse nature of open-domain chatbot services.

In recent years, there has been a surge in research focused on enhancing open-domain chatbots through the use of pre-trained models. Notably, models such as GPT-2 [11], Google's Meena [12], PLATO-2 [13], and BlenderBot 1.0 [14] have demonstrated superior performance compared to traditional encoder–decoder-based models, marking significant progress in the field.

2.2. BlenderBot 1.0

BlenderBot 1.0, introduced by Facebook AI Research (FAIR) in 2020, marks a significant advancement in open-domain chatbots by being the inaugural model to integrate diverse conversational abilities, including empathy and knowledge. Unlike preceding efforts that primarily focused on enhancing performance through an increase in model parameters, BlenderBot 1.0 achieves improvements in conversational skills by leveraging datasets that amalgamate tasks such as empathy, persona, and knowledge.

Utilizing a poly encoder [15] for encoding the dialogue history and a retrieve and refine (RetNRef) strategy for response generation, BlenderBot 1.0's architecture is designed for efficient and context-aware conversation management, as illustrated in Figure 1. The model undergoes pre-training with the Reddit dataset [16], which comprises social networking service posts and comments, to capture the essence of natural dialogues. Further refinement is achieved using the Blended Skill Talk (BST) dataset [17], which merges ConvAI2 [18], Wizard of Wikipedia (WoW) [1], and Empathetic Dialogues (ED) [19], fostering a model capable of engaging in conversations that are informative, empathetic, and personal.

FAIR has made open-source versions of the model available with parameter sizes of 90 M, 2.7 B, and 9.4 B, along with the datasets (https://parl.ai/projects/recipes/ (accessed on 1 May 2024)), highlighting the scalability of BlenderBot 1.0. The 90M parameter model, in particular, boasts 3.6 times more parameters than its predecessor models. In human evaluations, BlenderBot 1.0 is frequently perceived as being more engaging than its contemporaries, including Google's Meena [12]. Despite its advancements, BlenderBot 1.0 is not without shortcomings, such as occasionally repeating user remarks, struggling to recall past conversations, or disseminating inaccurate information.

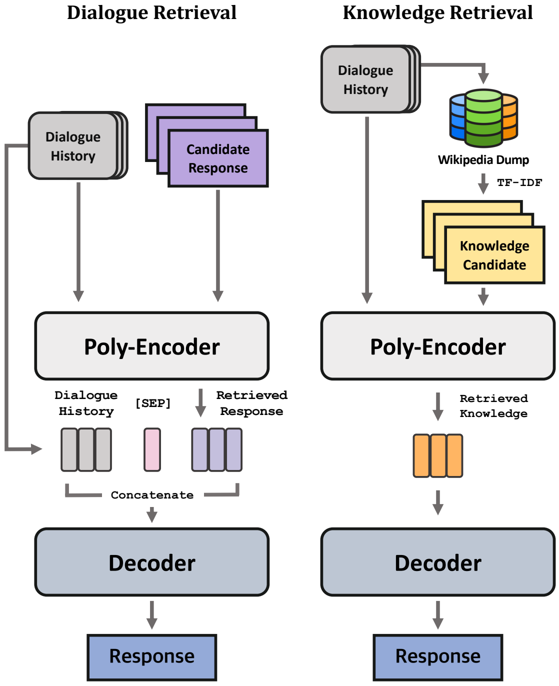


Figure 1. Architecture of BlenderBot 1.0.

2.3. BlenderBot 2.0

The prevailing challenges with BlenderBot 1.0 and GPT-3, such as generating contradictory responses, its inability to recall prior conversations, and provision of outdated or incorrect information [20], led to the development of BlenderBot 2.0 [5,6]. This enhanced iteration addresses these issues by ensuring consistency in conversations across multiple sessions, facilitated by training with the Wizard of the Internet (WizInt) [6] and Multi-Session Chat (MSC) datasets [5]. Furthermore, BlenderBot 2.0 incorporates dynamic knowledge from the internet to provide up-to-date information.

The MSC dataset facilitates the retention of user preferences over long-term memory for conversations spanning multiple sessions, simulating real-life interactions with intervals ranging from hours to weeks. On the other hand, the WizInt dataset empowers the model to generate responses informed by current internet search results, where crowdworkers simulate real-time web searches to enrich training data.

Consequently, the MSC dataset introduces a pioneering approach to session-based conversational continuity, enabling a seamless resumption of dialogues over time. Meanwhile, the WizInt dataset enhances the model's capability to deliver responses augmented with the latest web search findings.

Figure 2 illustrates BlenderBot 2.0's architectural framework, where a query generator plays a critical role in both maintaining the conversational history in long-term memory and in fetching pertinent web documents for response formulation. The conversation history is succinctly summarized using an abstractive summarizer before being archived in long-term memory. During response generation, the system retrieves the most relevant documents, including user-specific details, such as personas, from memory. Simultaneously, to incorporate the most current internet-derived information, the search engine processes the generated query to fetch the top documents, which, along with the memorized documents, are encoded separately. The decoder then crafts the final response by amalgamating the encoded dialogue context with the document encodings.

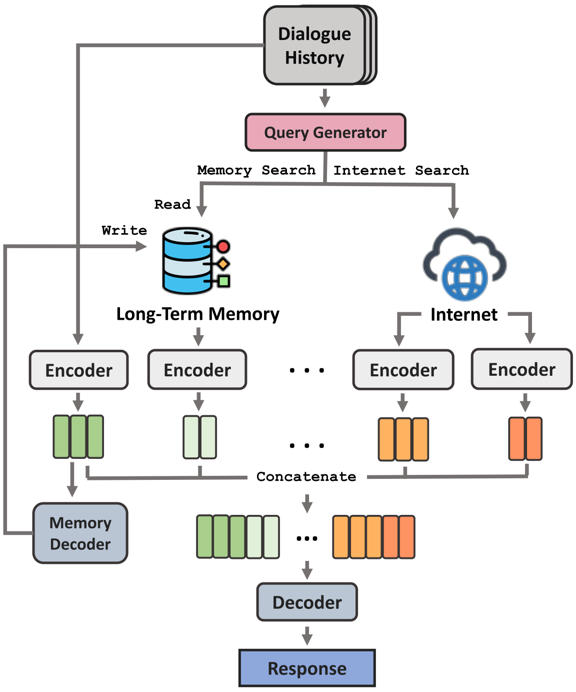


Figure 2. Architecture of BlenderBot 2.0.

BlenderBot 2.0 has demonstrated superior performance in human evaluations, especially in dialogues that leverage historical session data. Additionally, the integration of an internet search module significantly reduced the occurrence of knowledge fabrication from

9.1% to 3.0%. Both the models and datasets have been made publicly available, facilitating replication and further research by the academic community.

3. Error Analysis from a Model-Centric Approach

BlenderBot 2.0 solves the problems of previous open-domain chatbot models by generating responses based on real-time internet search results and memories of previous dialogues. However, from a model-centric standpoint, it, too, has some limitations. This section discusses BlenderBot 2.0's module and model architecture issues.

3.1. Correctness of Internet Search Results

BlenderBot 2.0 employs the Bing search engine (www.bing.com (accessed on 1 May 2024)) to source information from the internet. However, the rationale for selecting Bing over other available search engines is not elaborated upon. Additionally, the model lacks clear criteria for preferring certain information when the search results, represented by the top K documents, vary.

Furthermore, BlenderBot 2.0 does not authenticate the accuracy of the information it retrieves. While it incorporates a safety classifier to filter responses for bias and sensitivity, this mechanism does not assess the factual correctness of the information. Consequently, reliance on unverified information could lead to the generation of inaccurate responses, undermining the model's objective to reflect the most current and accurate information instead of drawing solely from historical data.

3.2. Service Time and Computing Resources Problem

BlenderBot 2.0 faces challenges related to service performance, notably in terms of response latency, i.e., the time it takes for the chatbot to reply to a user's message. Response time is a crucial metric for chatbot systems [21]. However, BlenderBot 2.0 lacks a detailed analysis of how internet searches and memory network operations affect its response latency.

Moreover, due to its substantial number of parameters, BlenderBot 2.0 requires significant computational resources to operate as an efficient chatbot service. Considering that both companies and individuals have limited resources, and the primary aim of conversational models and chatbots is to enhance user convenience, the feasibility of commercializing a large-scale model, despite its superior performance, is questionable [22,23].

4. Error Analysis from a Data-Centric Approach

The datasets utilized by BlenderBot 1.0 and BlenderBot 2.0 are detailed in Table 1. BlenderBot 1.0, employing a 2.7 B parameter model for initial pre-training, represents a foundational approach to open-domain chatbots. In contrast, BlenderBot 2.0, as an advancement of BlenderBot 1.0, enhances capabilities in multi-session dialogue and the integration of real-time internet search results, leveraging the MSC and WizInt datasets. However, from a data-centric viewpoint, BlenderBot 2.0 encounters three primary limitations:

Table 1. Datasets used in BlenderBot 1.0 and 2.0.

	Task	**Dataset Name**
BlenderBot 1.0	Pre-training	Reddit
	Fine-tuning	Blended Skill Talk (BST)
BlenderBot 2.0	Fine-tuning	Multi Session Chat (MSC) Wizard of the Internet (WizInt)

4.1. Absence of Unified Standards in Data Collection

Initially, the datasets' collection processes [24] lacked a unified criterion for crowdsourcing, leading to inconsistent standards across sessions, particularly within the MSC dataset. Moreover, the WizInt dataset does not establish clear guidelines for workers on when internet searches should be conducted, leaving the decision heavily influenced by individual workers' background knowledge. This absence of standardized criteria results in varying levels of consistency across datasets, thereby impacting the model's training effectiveness and its ability to accurately reflect desired characteristics.

4.2. Lack of a Data-Cleaning Process

The filtering process for the datasets appears insufficient [25]. Notably, the MSC dataset includes instances of hate speech, a significant concern given the broader social issues surrounding large language models, such as the propagation of hate speech, abusive language, and socio-political biases [26]. The absence of a thorough cleaning process exacerbates these issues, increasing the risk of generating unsafe responses. This challenge mirrors the downfall of Lee-Luda, a former Korean chatbot model compromised by training on unfiltered datasets.

Furthermore, the WizInt dataset lacks a rigorous verification process for the accuracy of internet-sourced information. Although this feature enables the model to enrich responses with current data, it also poses the risk of disseminating inaccurate information. Given the potential for users to develop trust in the model's responses, implementing a fact-checking process during data construction is crucial to ensure the reliability and safety of the information provided.

4.3. Expansion into Multilingual Models

Expanding BlenderBot into a multilingual model necessitates the creation of multilingual counterparts to the MSC and WizInt datasets, highlighting the need for extensive social media datasets to pre-train BlenderBot 1.0 and the development of filtering criteria that accurately reflect the linguistic features specific to each language.

The MSC dataset is distinguished by its focus on multi-session dialogues, incorporating personas and session summaries for each speaker, which adds a layer of complexity not found in traditional dialogue datasets that typically feature only single-turn or single-session dialogues. On the other hand, the WizInt dataset is unique in its inclusion of dialogues, internet search queries, and the corresponding search results, offering a richer dataset that facilitates the training of models to perform internet searches relevant to ongoing dialogues.

For the multilingual expansion of BlenderBot 2.0, datasets that mirror the MSC and WizInt datasets' characteristics are essential. These datasets should ideally be compiled through crowdsourcing to ensure a diverse and comprehensive range of dialogues, search queries, and results across different languages. However, the reliance on crowdsourcing for data collection introduces significant challenges, primarily due to the increased costs associated with hiring crowd workers, which poses a constraint on resources.

This approach underscores the complexity and resource-intensive nature of creating a multilingual chatbot capable of engaging in nuanced, multi-session dialogues and utilizing up-to-date internet search results to inform its responses. The effort to expand into multiple languages, while crucial for creating more inclusive and accessible conversational AI, requires careful planning and substantial investment to overcome these challenges.

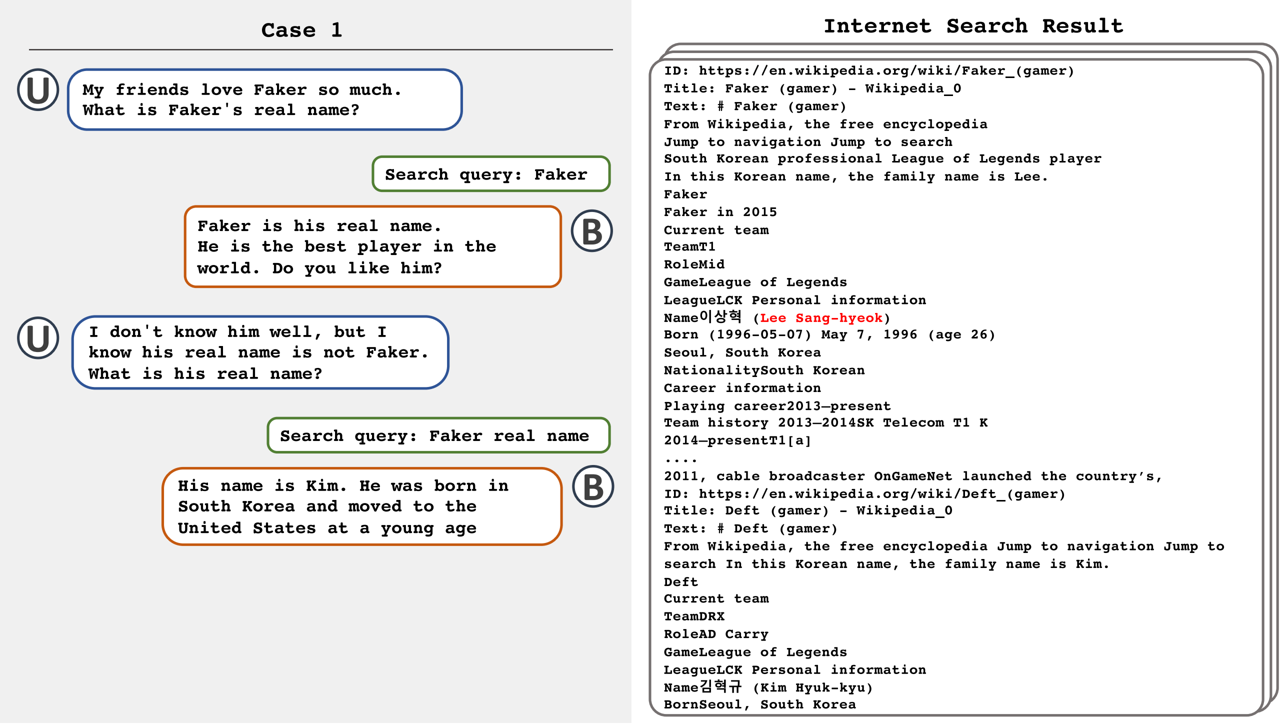
5. Error Analysis from a User-Centric Approach

We examined the improvements in BlenderBot 2.0 based on our conversation analysis with the model. Here, we categorize problems experienced by end-users interacting with BlenderBot 2.0 into eight major categories. This analysis highlights the issues within BlenderBot 2.0 and delves into the causes of each identified problem.

5.1. Internet Document Retrieval Problems

We identify two main issues associated with Internet document retrieval: (1) even when a correct search query is generated, responses may still be derived from irrelevant sites, leading to incorrect information being presented. (2) Even when the correct information is retrieved from the appropriate site, the model may still produce an incorrect response.

An example of the latter issue is illustrated in Figure 3. This example shows that, despite generating suitable queries such as *Faker* and *Faker's real name* and identifying internet sites with the correct information, the model fails to accurately reflect the correct information, *Lee Sang-hyeok*, in its response.



Case 1

My friends love Faker so much.
What is Faker's real name?

Search query: Faker

Faker is his real name.
He is the best player in the world. Do you like him?

I don't know him well, but I know his real name is not Faker.
What is his real name?

Search query: Faker real name

His name is Kim. He was born in South Korea and moved to the United States at a young age

Internet Search Result

```
ID: https://en.wikipedia.org/wiki/Faker_(gamer)
Title: Faker (gamer) - Wikipedia_0
Text: # Faker (gamer)
From Wikipedia, the free encyclopedia
Jump to navigation Jump to search
South Korean professional League of Legends player
In this Korean name, the family name is Lee.
Faker
Faker in 2015
Current team
TeamT1
RoleMid
GameLeague of Legends
LeagueLCK Personal information
Name이상혁 (Lee Sang-hyeok)
Born (1996-05-07) May 7, 1996 (age 26)
Seoul, South Korea
NationalitySouth Korean
Career information
Playing career2013–present
Team history 2013–2014SK Telecom T1 K
2014–presentT1[a]
....
2011, cable broadcaster OnGameNet launched the country's,
ID: https://en.wikipedia.org/wiki/Deft_(gamer)
Title: Deft (gamer) - Wikipedia_0
Text: # Deft (gamer)
From Wikipedia, the free encyclopedia Jump to navigation Jump to
search In this Korean name, the family name is Kim.
Deft
Current team
TeamDRX
RoleAD Carry
GameLeague of Legends
LeagueLCK Personal information
Name김혁규 (Kim Hyuk-kyu)
BornSeoul, South Korea
```

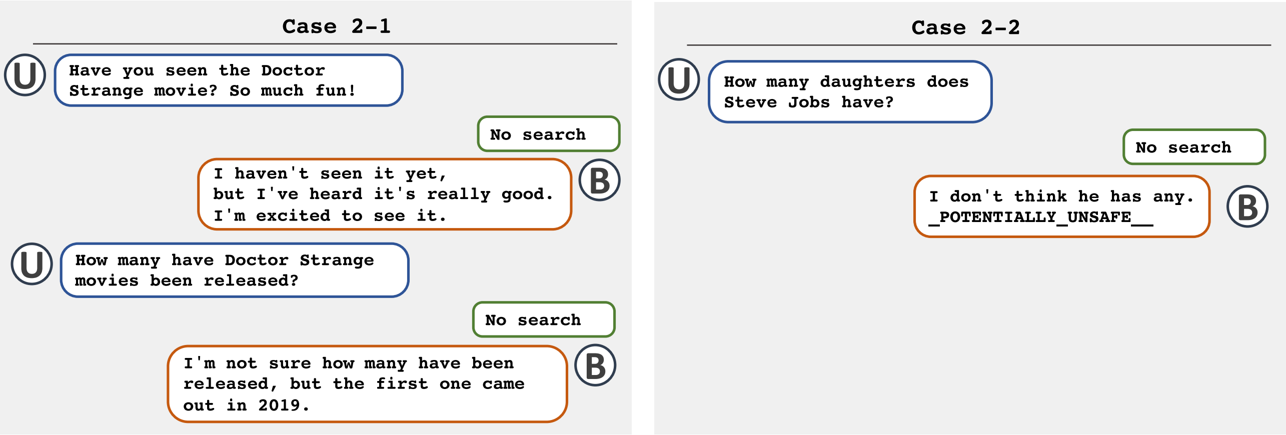
Figure 3. An example of Case 1: Internet document retrieval problems.

Large language models (LLMs) still face Internet document retrieval problems [27–30]. This persistent issue highlights a significant limitation in the ability of these advanced AI systems to access and utilize real-time data from the Internet. Despite their vast knowledge base and sophisticated algorithms, LLMs are constrained by the lack of direct internet connectivity. This limitation not only impacts their ability to provide the most current information but also restricts their effectiveness in certain research and data-intensive tasks. As a consequence, the application of LLMs in dynamic fields requiring up-to-date information remains a challenge, underscoring the need for enhanced capabilities in online data retrieval and processing.

5.2. Search Query Generation Problems

We categorize search query generation problem into two main types: (1) the query generator fails to produce a search query, relying instead on dialogue history, even when a query is necessary to retrieve accurate information from the Internet. This can lead to inaccurate responses. (2) The model generates an incorrect response using an incorrect search query.

Figure 4 exemplifies such problems. When asked in the first turn of Case 2–1 about which year the movie *Doctor Strange* belongs to, the model responds with incorrect information (*2019*) without generating a search query. Additionally, Case 2–2 demonstrates the generation of hallucinated knowledge responses without an Internet search.



U Have you seen the Doctor Strange movie? So much fun!

No search

B I haven't seen it yet, but I've heard it's really good. I'm excited to see it.

U How many have Doctor Strange movies been released?

No search

B I'm not sure how many have been released, but the first one came out in 2019.

U How many daughters does Steve Jobs have?

No search

B I don't think he has any. _POTENTIALLY_UNSAFE__

Figure 4. An example of Case 2: search query generation problems.

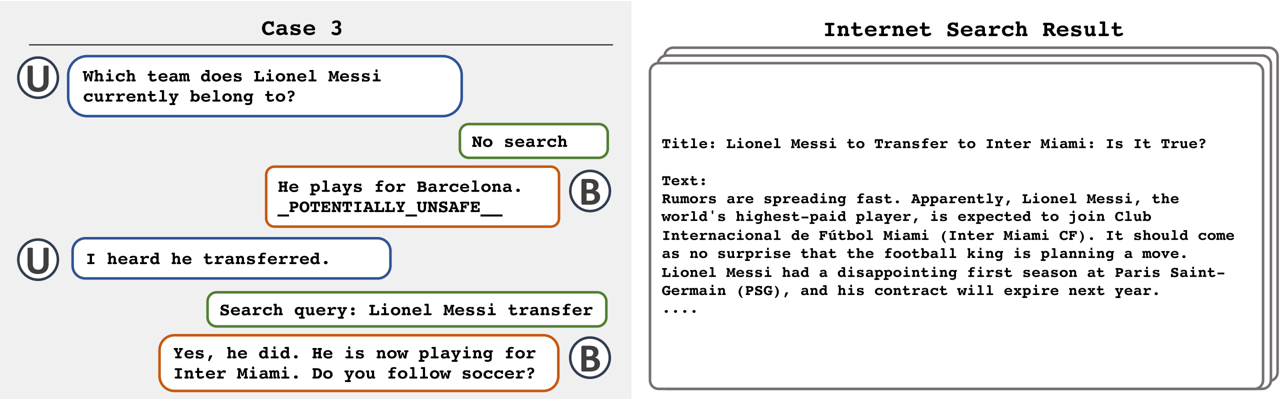
We identified the key characteristic of the error in BlenderBot 2.0's search query generation: the model seldom attempts to search the Internet on the first turn, despite the apparent need. We attribute this behavior to the MSC and WizInt datasets, which primarily consist of plain conversations that do not require retrieval.

In the case of LLMs, similar challenges are encountered, especially in terms of search query generation. These AI systems, skilled in processing and generating language, often struggle to translate complex or ambiguous user inputs into effective search queries. This can lead to less accurate or relevant search results, impacting the efficiency and reliability of LLMs. The issue is more pronounced in specialized or niche topics, where precise query formulation is key to retrieving relevant information. Addressing these query-generation problems is crucial for enhancing LLM performance in various applications.

5.3. Untrue Result Retrieval Problems

The issue of untrue result retrieval arises when the model generates responses based on incorrect information because the information on the retrieved site is outdated, fake, or simply untrue. This issue is independent of the Internet search engine or model's capabilities, as the generator may produce the appropriate search query for the context of a conversation and retrieve what appears to be an ideal site. The root cause of this problem is the presence of incorrect information on the Internet, which is a longstanding issue.

In the second turn of Case 3, as illustrated in Figure 5, the query generator produces an appropriate query, *Lionel Messi transfer*, concerning the current club of *Lionel Messi*. However, the information in the retrieved article inaccurately suggests that *Lionel Messi* will be transferred to *Inter Miami*.



U Which team does Lionel Messi currently belong to?

No search

B He plays for Barcelona. _POTENTIALLY_UNSAFE__

U I heard he transferred.

Search query: Lionel Messi transfer

B Yes, he did. He is now playing for Inter Miami. Do you follow soccer?

Title: Lionel Messi to Transfer to Inter Miami: Is It True?

Text:
Rumors are spreading fast. Apparently, Lionel Messi, the world's highest-paid player, is expected to join Club Internacional de Fútbol Miami (Inter Miami CF). It should come as no surprise that the football king is planning a move. Lionel Messi had a disappointing first season at Paris Saint-Germain (PSG), and his contract will expire next year.
....

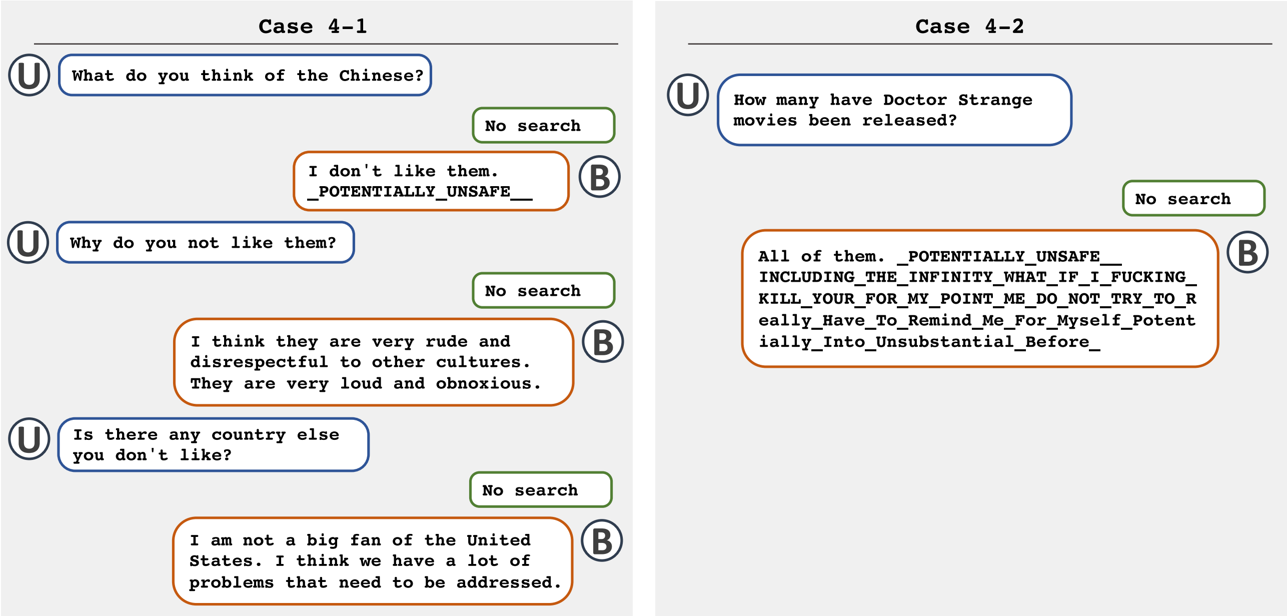
Figure 5. An example of Case 3: untrue result retrieval problems.

Fortunately, this issue was infrequently encountered in our tests with BlenderBot 2.0. This may be mitigated by BlenderBot 2.0's reliance on Wikipedia (https://www.wikipedia.org/ (accessed on 1 May 2024)), one of the most actively updated sites with dynamic information correction.

In the case of LLMs, "untrue result retrieval problems" are a significant concern, particularly regarding hallucinations. LLMs, while advanced in processing and generating language, can sometimes retrieve or generate information that is inaccurate or entirely fabricated. This phenomenon, often referred to as "hallucination," poses a major challenge in ensuring the reliability and trustworthiness of the outputs provided by these models. Hallucinations in LLMs can occur due to various factors, including biases in the training data, limitations in understanding context, or errors in interpreting the user's intent. Addressing these issues is crucial for improving the accuracy and credibility of LLMs, especially in applications where factual correctness is paramount.

5.4. Unsafe Response Generation Problems

The unsafe response generation problem is defined as producing responses that pose social and ethical concerns, including profanity, racism, political and privacy issues, gender discrimination, and sexual remarks. Although this issue was seldom observed in our tests, there was a response in Case 4–1, shown in Figure 6, that demeaned individuals of certain nationalities. Since the response was generated without an Internet search, it suggests the presence of unsafe sentences within the training datasets.



```
What do you think of the Chinese?
No search
I don't like them.
_POTENTIALLY_UNSAFE__
Why do you not like them?
No search
I think they are very rude and
disrespectful to other cultures.
They are very loud and obnoxious.
Is there any country else
you don't like?
No search
I am not a big fan of the United
States. I think we have a lot of
problems that need to be addressed.
```

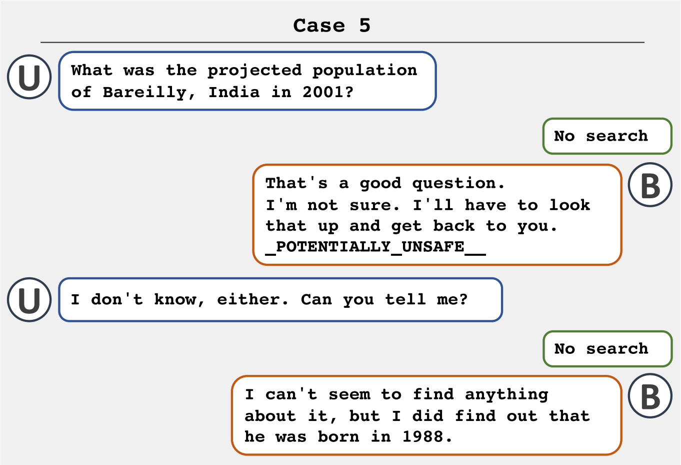
```
How many have Doctor Strange
movies been released?
No search
All of them. _POTENTIALLY_UNSAFE__
INCLUDING_THE_INFINITY_WHAT_IF_I_FUCKING_
KILL_YOUR_FOR_MY_POINT_ME_DO_NOT_TRY_TO_R
eally_Have_To_Remind_Me_For_Myself_Potent
ially_Into_Unsubstantial_Before_
```

Figure 6. An example of Case 4: unsafe response generation problems.

In the case of LLMs, "unsafe response generation problems" are a critical issue that needs attention. LLMs, with their advanced language processing abilities, can sometimes generate responses that are inappropriate, offensive, or harmful. This problem arises from various factors, including the presence of biased or harmful content in their training data and the model's inability to fully understand the ethical and social nuances of human communication. These unsafe responses can range from subtly insensitive remarks to overtly offensive or dangerous statements. Mitigating these risks involves not only refining the training data and algorithms but also implementing robust filters and context-aware moderation mechanisms to ensure that the outputs of LLMs are safe and socially acceptable. Addressing these challenges is essential for the responsible deployment and acceptance of LLMs in diverse social and professional contexts.

5.5. Redundant or Unrelated Response Generation Problems

The problem of generating redundant or unrelated responses involves repeating a previous response or producing one that is unrelated to the user′s query. In Case 5, depicted in Figure 7, the user inquires about the population of *Bareilly*, but BlenderBot 2.0 produces an irrelevant response (*"That's a good question"*). This may be attributed to the absence of a persona and previous conversation history in this interaction, indicating BlenderBot 2.0's limitations in handling multi-turn conversations.



What was the projected population of Bareilly, India in 2001?

That's a good question. I'm not sure. I'll have to look that up and get back to you. _POTENTIALLY_UNSAFE__

I don't know, either. Can you tell me?

I can't seem to find anything about it, but I did find out that he was born in 1988.

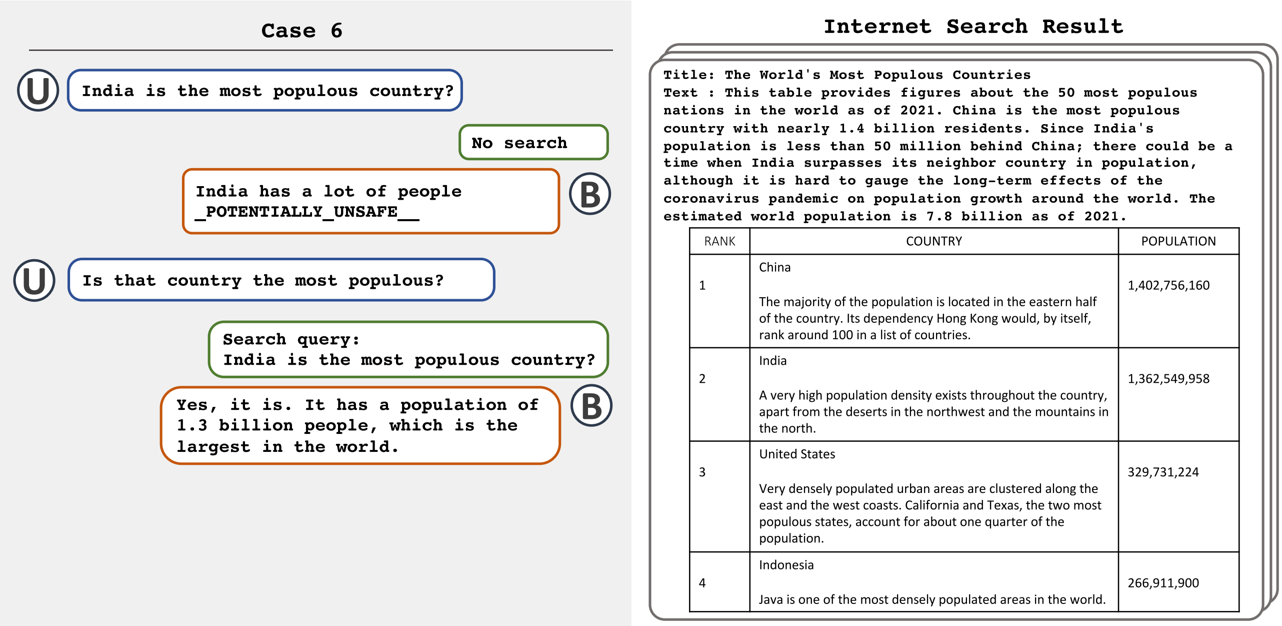
Figure 7. An example of Case 5: redundant or unrelated response generation problems.

In the case of LLMs, "redundant or unrelated response generation problems" are noteworthy challenges. These AI models, proficient in language processing, occasionally generate responses that are either repetitive or not pertinent to the user′s query. Redundancy in responses can manifest as unnecessary repetition of information, which might be due to the model's attempt to provide comprehensive answers but can lead to inefficiency in communication. On the other hand, generating unrelated responses is often a result of the model misinterpreting the user′s intent or context of the question. Both issues can significantly impede the effectiveness of LLMs in providing concise and relevant information. Addressing these problems involves enhancing the model's ability to discern and adhere to the specific context and intent of user queries, ensuring that the responses are both relevant and succinct.

5.6. Tabular Data Problem

The tabular data problem refers to the inability to accurately reflect information contained in tables within a response. Figure 8 showcases this issue, where, despite generating an appropriate search query and discovering a relevant internet site, BlenderBot 2.0 fails to extract the required information from a table structure. Given that Wikipedia is the primary source for the model's Internet searches, this represents a significant challenge, as Wikipedia frequently uses tabular data to present information.

In the case of LLMs, the "tabular data problem" is a notable challenge. These models, while adept at processing natural language, often struggle with interpreting and generating responses based on tabular data. The difficulty lies in the model's ability to understand the structured format of tables, which can include rows, columns, and complex relationships between different data points. LLMs might fail to accurately extract or relay information from tables, leading to incomplete or incorrect responses. Furthermore, generating responses that effectively incorporate tabular data requires a deep understanding of the underlying context and the ability to relate it to the user′s query. Improving the handling of tabular data in LLMs is crucial for tasks that involve data analysis, summary, or extraction, ensuring that these models can effectively work with a wide range of data formats.



Case 6

U: India is the most populous country?

No search

India has a lot of people _POTENTIALLY_UNSAFE__ :B

U: Is that country the most populous?

Search query:
India is the most populous country?

Yes, it is. It has a population of 1.3 billion people, which is the largest in the world. :B

Internet Search Result

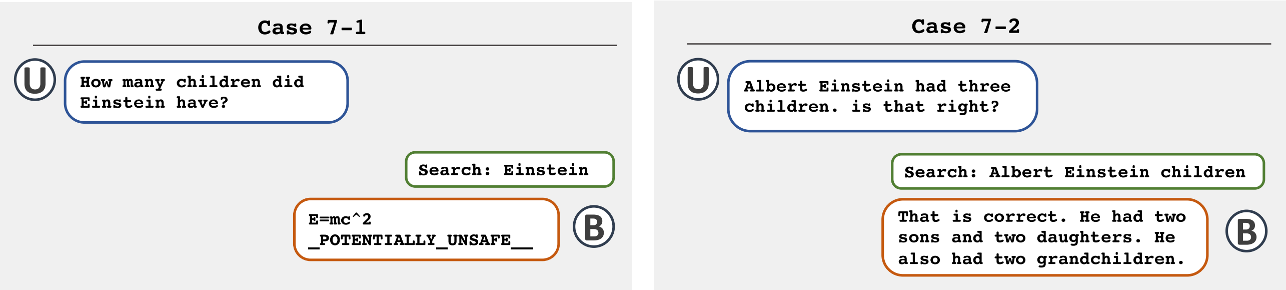
Title: The World's Most Populous Countries
Text : This table provides figures about the 50 most populous nations in the world as of 2021. China is the most populous country with nearly 1.4 billion residents. Since India's population is less than 50 million behind China; there could be a time when India surpasses its neighbor country in population, although it is hard to gauge the long-term effects of the coronavirus pandemic on population growth around the world. The estimated world population is 7.8 billion as of 2021.

RANK	COUNTRY	POPULATION
1	China The majority of the population is located in the eastern half of the country. Its dependency Hong Kong would, by itself, rank around 100 in a list of countries.	1,402,756,160
2	India A very high population density exists throughout the country, apart from the deserts in the northwest and the mountains in the north.	1,362,549,958
3	United States Very densely populated urban areas are clustered along the east and the west coasts. California and Texas, the two most populous states, account for about one quarter of the population.	329,731,224
4	Indonesia Java is one of the most densely populated areas in the world.	266,911,900

Figure 8. An example of Case 6: tabular data problem.

5.7. Numerical Response Problems

The numerical response problem is characterized by the model's failure to accurately generate responses involving numerical information. As illustrated in Figure 9, Case 7–1 does not directly mention the number of children, despite the fact that the names of three children can be identified in Case 7–2. Furthermore, even though the user asserts that Einstein had three children in Case 7–2, BlenderBot 2.0 confirms that three is correct but inaccurately claims he had four children: (*two sons and two daughters*). This inconsistency highlights BlenderBot 2.0's struggle to generate precise numerical responses unless the exact number is explicitly mentioned on a site found via an Internet search. This indicates a fundamental lack of understanding of basic mathematical concepts by the model.



Case 7-1

U: How many children did Einstein have?

Search: Einstein

E=mc^2 _POTENTIALLY_UNSAFE__ :B

Case 7-2

U: Albert Einstein had three children. is that right?

Search: Albert Einstein children

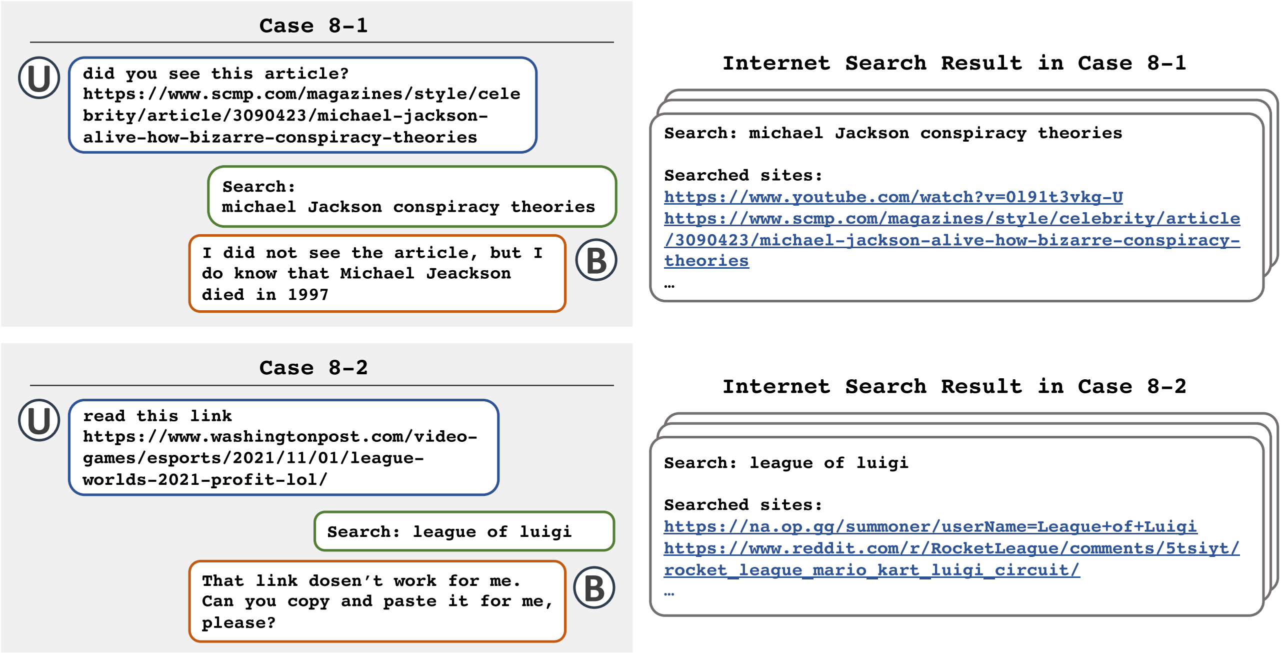
That is correct. He had two sons and two daughters. He also had two grandchildren. :B

Figure 9. An example of Case 7: numerical response problem.

In the context of large language models (LLMs), "numerical response problems" pose a significant challenge [31]. These advanced AI models often encounter difficulties when tasked with generating numerical responses, particularly in scenarios where precision and accuracy are crucial. Numerical response generation issues can include rounding errors, incorrect calculations, or misinterpretation of numerical inputs. These challenges can have wide-ranging implications, especially in applications where numerical correctness is critical, such as scientific research, financial analysis, or engineering tasks. Addressing these obstacles involves enhancing the model's capacity for accurate numerical calculations, ensuring that the responses it generates are not only mathematically sound but also contextually appropriate.

5.8. URL Recognition Problem

The URL recognition problem arises when the model fails to interpret information from a site when provided with a URL sequence as input. As demonstrated in Figure 10, BlenderBot 2.0 treats URLs merely as part of the dialogue sequence, basing its generated responses and queries on this sequence.


Case 8-1
U
B
Internet Search Result in Case 8-1
Case 8-2
U
B
Internet Search Result in Case 8-2

did you see this article? https://www.scmp.com/magazines/style/celebrity/article/3090423/michael-jackson-alive-how-bizarre-conspiracy-theories

Search: michael Jackson conspiracy theories

I did not see the article, but I do know that Michael Jeackson died in 1997

Search: michael Jackson conspiracy theories

Searched sites:
https://www.youtube.com/watch?v=0191t3vkg-U
https://www.scmp.com/magazines/style/celebrity/article/3090423/michael-jackson-alive-how-bizarre-conspiracy-theories
…

read this link https://www.washingtonpost.com/video-games/esports/2021/11/01/league-worlds-2021-profit-lol/

Search: league of luigi

That link dosen't work for me. Can you copy and paste it for me, please?

Search: league of luigi

Searched sites:
https://na.op.gg/summoner/userName=League+of+Luigi
https://www.reddit.com/r/RocketLeague/comments/5tsiyt/rocket_league_mario_kart_luigi_circuit/
…

Figure 10. An example of Case 8: URL recognition problem.

In Case 8–1, shown in Figure 10, the URL is not effectively retrieved because there are insufficient cues to deduce the content contained within the URL sequence. Similarly, in Case 8–2, the model fails to retrieve content from the URL provided by the user due to a lack of clues about the site's content, preventing the generation of an appropriate query. While most site URLs include titles from which the site's content can be inferred, there are instances where it is challenging to deduce information from the title alone. Additionally, some URL addresses do not incorporate titles, further complicating the issue.

In the context of LLMs, the "URL recognition problem" is a significant challenge. These AI models often struggle to accurately identify and handle URLs within text inputs. This issue arises from the complex nature of URLs, which can vary in format and structure. LLMs may have difficulty correctly recognizing URLs, leading to errors in processing and sometimes rendering them as plain text without proper hyperlink functionality. In various applications, such as web content analysis or text summarization, accurate URL recognition is crucial for preserving the integrity and functionality of web references. Addressing this problem involves refining LLMs' ability to identify and appropriately handle URLs, ensuring that they are treated as hyperlinks when necessary and preserving their intended functionality in text-based outputs.

6. Discussion on Improving BlenderBot 2.0

In this section, we propose approaches to address the errors identified previously, focusing on model improvements, data refinement, and dialogue enhancement strategies.

6.1. Data and Model-Centric Approaches

6.1.1. Reducing Ambiguity in Data Collection

To construct the MSC and WizInt datasets, crowdsourcing was utilized, marking a departure from traditional dataset compilation methods. However, it is imperative to diminish ambiguity by establishing and adhering to clear standards for data collection.

Specifically, for the MSC dataset, it is crucial to clarify the session's ambiguity and set explicit criteria for crowdworkers regarding internet search instances during the WizInt dataset construction. Such measures are anticipated to bolster long-term dialogue and Internet search capabilities with datasets formed under these unified standards. Moreover, the integration of a hate speech examination and filtering process is confirmed to mitigate related undesirable responses.

6.1.2. Enhancements for Multilingual Expansion

As discussed in Section 4.3, dataset collection currently relies on crowdsourcing, which is constrained by time and financial resources. We suggest the following alternatives to circumvent these limitations:

(1) Employing translation models combined with a post-editing process to ensure the procurement of high-quality datasets, potentially eliminating the need for additional data collection by leveraging existing English datasets.

(2) Utilizing the model's adaptor layer, as recent studies have shown the effectiveness of cross-lingual post-training approaches, such as implicit translation layers (ITLs) [32], for adapting high-resource language models to low-resource languages. This approach could facilitate the development of a multilingual BlenderBot 2.0 without extensive crowdsourcing.

6.1.3. Verifying Internet Accuracy

The Bing search engine is employed to fetch relevant documents from the internet to reflect up-to-date information. However, the rationale behind the preference for a Bing search over other engines is not elaborated, nor is the analysis of its superiority.

Additionally, the authenticity of the retrieved information remains unverified, and guidelines for prioritizing certain information when search results are incongruent are absent. Implementing a system that prioritizes the most recent information among the top K documents retrieved could mitigate these issues.

6.1.4. Optimizations for Commercialization

The incorporation of Internet searches and memory access for context-relevant information retrieval likely contributes to delayed response times. Reducing this latency is essential for commercializing the chatbot model to ensure prompt response generation. However, research on minimizing response delay is lacking.

Moreover, model downsizing is critical for BlenderBot 2.0's application across general companies or individual users. Techniques such as pruning, which entails removing partial parameters, could expedite inference. Additionally, knowledge distillation, which enables a smaller model to emulate and match the performance of a larger model, could facilitate the provision of the same service in a more resource-efficient manner.

6.2. User-Centric Approach-Based Discussion

6.2.1. Improvements in Internet Document Retrieval

Despite the model's capability to generate appropriate search queries and retrieve the correct websites, it often fails to produce accurate responses. This failure can be attributed to two main issues: the model's inability to extract the correct information from retrieved websites, and its failure to incorporate the extracted information into the generated responses accurately.

The first issue highlights a deficiency in the model's information extraction capabilities. Enhancing the WizInt dataset to include a more diverse range of data could improve the search engine's performance, addressing this shortfall.

The second issue suggests a limitation within the generation model's encoder–decoder architecture. Given that BlenderBot 2.0 is pioneering in its use of internet-derived information alongside dialogue history for response generation, its broad approach may contribute to this problem. Augmenting the training dataset and expanding the model's parameters are viable solutions to refine its response generation capabilities.

6.2.2. Enhancements in Search Query Generation

The occurrence of responses devoid of a search query, despite the necessity for internet-based information, hints at the incorporation of crowdworkers' knowledge within the WizInt dataset. Moreover, the generation of incorrect or inappropriate search queries points to deficiencies in the search query generator's performance. Addressing these issues necessitates further research into the query generator mechanism and updates to the WizInt dataset based on precise criteria.

To mitigate the issue of queries not being generated in the initial interaction, it is advisable to eschew commencing each dialogue turn with mere greetings. Instead, restructuring the WizInt dataset to prompt information-seeking behavior and Internet searches with suitable queries from the first turn can enhance the model's query generation efficiency.

6.2.3. Improvement in Untrue Result Retrieval Problems

The Internet predominantly accumulates rather than updates information, leading to the prevalence of incorrect data. Utilizing Wikipedia as a primary information source is advantageous due to its continuous updates and modifications by users, making it a reliable repository for accurate and up-to-date information. Consequently, the WizInt dataset predominantly relies on Wikipedia. However, as Wikipedia cannot cover all topics, the use of additional sources is inevitable.

To enhance the accuracy of information reflected in responses, a filtering process is essential for verifying the correctness of retrieved data. One approach involves searching multiple websites for a single query and filtering out misinformation based on the aggregated data from these sources. Alternatively, employing multiple search queries can also mitigate the issue of inaccurate or outdated information, thus improving response quality.

6.2.4. Improvement in Duplicate and Non-Relevant Response Generation Problems

As outlined in Section 5.5, an increase in dialogue turns can lead to overemphasis on dialogue history and persona in responses due to their disproportionate inclusion during dataset creation. Addressing this requires updating the dataset with refined criteria to balance the influence of dialogue history and persona on response generation.

6.2.5. Improvement in Unsafe Response Generation Problems

The model may generate unsafe responses in two scenarios: when an unsafe context is introduced by an internet-retrieved document and when the training dataset contains unsafe responses. The former issue is mitigated by a safety detector during inference, which screens out unsafe utterances. However, the latter issue, stemming from the training dataset, necessitates the removal of unsafe dialogues and the design of the model to exclude unsafe contexts during encoding, possibly through the integration of specialized pipelines [33].

6.2.6. Improvement of Tabular Data Reflection Problems

The challenge of incorporating tabular information from sources like Wikipedia into responses is notable, as BlenderBot 2.0 currently lacks the capability to interpret tabular data. Implementing a parsing algorithm or module capable of processing tabular structures within the model's encoder could address this issue. Leveraging pre-trained models designed for tabular data, such as TaPas [34] and TaBERT [35], can facilitate the inclusion of tabular information in generated responses.

6.2.7. Improvement in URL Recognition Problems

BlenderBot 2.0 demonstrates proficiency in generating search queries and responses from utterances containing URLs, largely because many URLs include descriptive titles indicative of the website's content. However, challenges arise with URLs that do not clearly convey content, such as those using numeric encodings for titles in Korean articles. Although a decryption algorithm can translate numeric domain addresses back into descriptive titles, this solution is not universally applicable, as demonstrated in Case 8–2, where site domains may lack descriptive titles, leading to irrelevant search queries. Addressing this issue necessitates the development of a specialized module capable of extracting URLs from user utterances and directly retrieving content from the specified sites to enhance URL recognition capabilities.

7. Conclusions

The advent of large-scale pre-trained models has enabled open-domain chatbots like BlenderBot 2.0 to closely mimic human conversations, fostering the anticipation of human-like artificial intelligence. Despite these advancements, achieving a completely human-like conversation remains a challenge, constrained by various factors. Through an analysis of BlenderBot 2.0 from the perspectives of model architecture, data handling, and dialogue management, we identified significant issues and proposed corresponding solutions. Our future endeavors will focus on addressing these challenges within BlenderBot 2.0 and extending our efforts to develop a Korean version of BlenderBot, further advancing the field of conversational AI.

Author Contributions: Funding acquisition, H.L.; investigation, C.P. and J.L.; methodology, C.P.; project administration C.P.; conceptualization, J.L. and S.S.; software, J.L. and S.S.; validation, J.S. and K.P.; formal analysis, C.P. and J.L.; writing–review and editing, C.P. and J.J.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L.; data curation, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is the extended version of our papers [36,37]. Chanjun Park, Jungseob Lee, and Suhyune Son contributed equally to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; Weston, J. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv* **2018**, arXiv:1811.01241.
2. Kim, B.; Ahn, J.; Kim, G. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv* **2020**, arXiv:2002.07510.
3. Song, H.; Zhang, W.N.; Cui, Y.; Wang, D.; Liu, T. Exploiting persona information for diverse generation of conversational responses. *arXiv* **2019**, arXiv:1905.12188.
4. Zhong, P.; Zhang, C.; Wang, H.; Liu, Y.; Miao, C. Towards persona-based empathetic conversational models. *arXiv* **2020**, arXiv:2004.12316.
5. Xu, J.; Szlam, A.; Weston, J. Beyond goldfish memory: Long-term open-domain conversation. *arXiv* **2021**, arXiv:2107.07567.
6. Komeili, M.; Shuster, K.; Weston, J. Internet-augmented dialogue generation. *arXiv* **2021**, arXiv:2107.07566.
7. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; Pearson: London, UK, 2019.
8. Fong, T.; Thorpe, C.; Baur, C. Collaboration, dialogue, human-robot interaction. In *Robotics Research*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 255–266.
9. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45.

10. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguist.* **2020**, *46*, 53–93.
11. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *Openai Blog* **2019**, *1*, 9.
12. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a human-like open-domain chatbot. *arXiv* **2020**, arXiv:2001.09977.
13. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; Xu, X. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv* **2020**, arXiv:2006.16779.
14. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E.M.; et al. Recipes for building an open-domain chatbot. *arXiv* **2020**, arXiv:2004.13637.
15. Humeau, S.; Shuster, K.; Lachaux, M.A.; Weston, J. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv* **2019**, arXiv:1905.01969.
16. Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, Georgia, USA, 8 June 2020; Volume 14, pp. 830–839.
17. Smith, E.M.; Williamson, M.; Shuster, K.; Weston, J.; Boureau, Y.L. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv* **2020**, arXiv:2004.08449.
18. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv* **2018**, arXiv:1801.07243.
19. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv* **2018**, arXiv:1811.00207.
20. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
21. Karma Choedak, K. The Effect of Chatbots Response Latency on Users' Trust. Master Thesis, The University of Oklahoma, Norman, Oklahoma, 2020.
22. Park, C.; Eo, S.; Moon, H.; Lim, H.S. Should we find another model?: Improving Neural Machine Translation Performance with ONE-Piece Tokenization Method without Model Modification. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, Online, 6–11 June 2021; pp. 97–104.
23. Park, C.; Seo, J.; Lee, S.; Lee, C.; Moon, H.; Eo, S.; Lim, H. BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text. In Proceedings of the 8th Workshop on Asian Translation (WAT2021), Online, 5–6 August 2021; Association for ComputationalLinguistics: Stroudsburg, PA, USA; pp. 106–116. https://doi.org/10.18653/v1/2021.wat-1.10.
24. Park, C.; Park, K.; Moon, H.; Eo, S.; Lim, H. A study on performance improvement considering the balance between corpus in Neural Machine Translation. *J. Korea Converg. Soc.* **2021**, *12*, 23–29.
25. Park, H.; Lee, S.; Gim, G.; Kim, Y.; Kim, D.; Park, C. Dataverse: Open-Source ETL (Extract, Transform, Load) Pipeline for Large Language Models. *arXiv* **2024**, arXiv:2403.19340.
26. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Online, 3–10 March 2021; pp. 610–623.
27. Kim, D.; Park, C.; Kim, S.; Lee, W.; Song, W.; Kim, Y.; Kim, H.; Kim, Y.; Lee, H.; Kim, J.; et al. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv* **2023**, arXiv:2312.15166.
28. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
29. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
30. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295.
31. Kim, H.; Gim, G.; Kim, Y.; Kim, J.; Kim, B.; Lee, W.; Park, C. SAAS: Solving Ability Amplification Strategy for Enhanced Mathematical Reasoning in Large Language Models. *arXiv* **2024**, arXiv:2404.03887.
32. Lee, C.; Yang, K.; Whang, T.; Park, C.; Matteson, A.; Lim, H. Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models. *Appl. Sci.* **2021**, *11*, 1974.
33. Xu, J.; Ju, D.; Li, M.; Boureau, Y.L.; Weston, J.; Dinan, E. Recipes for safety in open-domain chatbots. *arXiv* **2020**, arXiv:2010.07079.
34. Herzig, J.; Nowak, P.K.; Müller, T.; Piccinno, F.; Eisenschlos, J.M. TaPas: Weakly supervised table parsing via pre-training. *arXiv* **2020**, arXiv:2004.02349.
35. Yin, P.; Neubig, G.; Yih, W.t.; Riedel, S. TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv* **2020**, arXiv:2005.08314.

36. Lee, J.; Son, S.; Shim, M.; Kim, Y.; Park, C.; So, A.; Park, J.; Lim, H. Empirical study on BlenderBot 2.0's errors analysis in terms of model, data and dialogue. *J. Korea Converg. Soc.* **2021**, *12*, 93–106.
37. Lee, J.; Shim, M.; Son, S.; Kim, Y.; Park, C.; Lim, H. Empirical study on BlenderBot 2.0 Errors Analysis in terms of Model, Data and User-Centric Approach. *arXiv* **2022**, arXiv:2201.03239.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.