Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

General IS Topics

Dec 11th, 12:00 AM

A User-centric Taxonomy for Conversational Generative Language Models

Constantin von Brackel-Schmidt
Universität Hamburg, constantin.schmidt@uni-hamburg.de

Emir Kučević
Universität Hamburg, emir.kucevic@uni-hamburg.de

Lucas Memmert
Universität Hamburg, lucas.memmert@uni-hamburg.de

Navid Tavanapour
Universität Hamburg, navid.tavanapour@uni-hamburg.de

Izabel Cvetkovic
University of Hamburg, izabel.cvetkovic@uni-hamburg.de

See next page for additional authors

Recommended Citation

von Brackel-Schmidt, Constantin; Kučević, Emir; Memmert, Lucas; Tavanapour, Navid; Cvetkovic, Izabel; Bittner, Eva A. C.; and Böhmann, Tilo, "A User-centric Taxonomy for Conversational Generative Language Models" (2023). *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*. 1.
https://aisel.aisnet.org/icis2023/generalis/generalis/1

Presenter Information

Constantin von Brackel-Schmidt, Emir Kučević, Lucas Memmert, Navid Tavanapour, Izabel Cvetkovic, Eva A. C. Bittner, and Tilo Böhmann

A User-centric Taxonomy for Conversational Generative Language Models

Completed Research Paper

Constantin von Brackel-Schmidt
Universität Hamburg
Hamburg, Germany
constantin.schmidt@uni-hamburg.de

Emir Kučević
Universität Hamburg
Hamburg, Germany
emir.kucevic@uni-hamburg.de

Lucas Memmert
Universität Hamburg
Hamburg, Germany
lucas.memmert@uni-hamburg.de

Navid Tavanapour
Universität Hamburg
Hamburg, Germany
navid.tavanapour@uni-hamburg.de

Izabel Cvetkovic
Universität Hamburg
Hamburg, Germany
izabel.cvetkovic@uni-hamburg.de

Eva A. C. Bittner
Universität Hamburg
Hamburg, Germany
eva.bittner@uni-hamburg.de

Tilo Böhmann
Universität Hamburg
Hamburg, Germany
tilo.boehmann@uni-hamburg.de

Abstract

Conversational generative language models (GLMs) like ChatGPT are being rapidly adopted. Previous research on non-conversational GLMs showed that formulating prompts is critical for receiving good outputs. However, it is unclear how conversational GLMs are used when solving complex problems that require multi-step interactions. This paper addresses this research gap based on findings from a large participant event we conducted, where ChatGPT was iteratively and in a multi-step manner used while solving a complex problem. We derived a taxonomy of prompting behavior employed for solving complex problems as well as archetypes. While the taxonomy provides common knowledge on GLMs usage based on analyzed input-prompts, the different archetypes facilitate the classification of operators according to their usage. With both we provide exploratory knowledge and a foundation for design science research endeavors, which can be referred to, enabling further research and development of prompt engineering, prompting tactics, and prompting strategies on common ground.

Keywords: Generative Language Models, Generative AI, human-computer-interaction

Introduction

Advancements in artificial intelligence (AI) have led to increased interest in how AI can be leveraged to support complex problem solving (Krogh, 2018), moving towards humans and AI systems solving problems collaboratively (Akata et al., 2020; Dellermann et al., 2019). While addressing open-ended problems might have been difficult with more traditional AI approaches (Krogh, 2018), generative language models (GLMs) as a new type of machine learning model have been shown to be useful in supporting humans in almost unbounded area of language-mediated tasks (Brown et al., 2020; Lester et al., 2021; Mishra, Khashabi, Baral, & Hajishirzi, 2022; Swanson et al., 2021; Wu et al., 2022). However, it is still an open question how these models can best be used for different tasks.

With the recent introduction of a chat-based interaction mode, the generative capabilities of GLMs (e.g., OpenAI's ChatGPT) have become broadly available, even to non-technical users (Jiang et al., 2022). Models such as OpenAI's GPT perform well at generating free text based on a given input, even without additional task-specific training (Brown et al., 2020). Since GLM system output is conditioned on input during inference time (Brown et al., 2020), it might be challenging to create a suitable formulation for the input request to receive an expected generated output (Jiang et al., 2022), reflecting the research field related to prompt engineering (Lester et al., 2021; Mishra, Khashabi, Baral, Choi, et al., 2022). Studies on this topic already investigate methods to tune initial input-prompts to generate more optimal outputs for solving complex problems with GLM systems (Devlin et al., 2019; Hambardzumyan et al., 2021; Jiang et al., 2022; Lester et al., 2021; Liu et al., 2023; Mishra, Khashabi, Baral, Choi, et al., 2022). Scholars have shown that even slight modifications of initial prompts might result in different GLM system outputs (Jiang et al., 2022; Lester et al., 2021; Liu et al., 2023; Mishra, Khashabi, Baral, Choi, et al., 2022). While outputs to initial prompts might produce unexpected content, GLM system users seem to develop tactics for working with such systems to solve complex tasks, such as task allocation, in prompt chains, if provided with building blocks (Wu et al., 2022).

The release of the chat-based interaction mode with GLM enables long conversations (Jiang et al., 2022), and it might open new ways for tactics to work with such systems to solve complex problems. It is unclear whether and how the chat-based interaction mode might affect known prompting tactics or lead to new prompting strategies for users. Therefore, more research in this field is needed. Additionally, limited knowledge exists about how humans use GLM systems (Jiang et al., 2022; Lester et al., 2021; Mishra, Khashabi, Baral, Choi, et al., 2022; Wu et al., 2022) to solve complex problems (Akata et al., 2020; Dellermann et al., 2019), especially if they aim to generate specific content (Jiang et al., 2022; Mishra, Khashabi, Baral, Choi, et al., 2022; Wu et al., 2022). We addressed this research gap by focusing the present study on the following research question (RQ):

RQ: On an operational level, how do humans prompt GLM-based systems to attempt solving complex problems?

To address the RQ, we investigated how humans interact with GLM systems to solve a complex problem with a novel format for human-AI collaboration. We designed one of the first Prompt-a-thon sessions, in which a diverse set of 76 participants worked on specific challenges in small groups with the support of GLMs. Based on the logs of these interactions, we conducted a qualitative content analysis of the input-prompts as a foundation for developing an extendable taxonomy in accordance with Nickerson et al. (2013) and Kundisch et al. (2021b) to characterize human prompting approaches for GLMs. We contribute a taxonomy that offers common ground of conversational GLMs usage conducing researchers and practitioners. Practitioners can use it when utilizing GLMs for specific tasks such as complex problem solving or relate to it for prompt engineering. Researchers can use it to classify and analyze prompts in multi-step interactions, facilitating research for example on prompt strategies. It represents the first comprehensive terminology on GLM usage and its prompts in this field, setting the stage for subsequent research and the evolution of pertinent theories, as noted by Kundisch et al. (2021b). Additionally, we derived archetypes from the collected data to recognize prompting patterns of humans when interacting with GLMs. Overall, our results provide first insights into prompting approaches during conversations.

The remainder of this paper is structured as follows. In the next section, we review related research on GLMs and prompt engineering. Then, we describe our research methodology and how we collected and analyzed the data to develop the taxonomy and derive the archetypes. Subsequently, we present and discuss our results before highlighting the theoretical and practical contributions of our research. After that, we mention the limitations of our research and provide suggestions for future research before closing the paper with a conclusion.

Related Research

Large language models are a specific type of machine learning-based models trained on large corpora of texts (Brown et al., 2020). Such models are trained to generate the next word(s) upon input, consequently referred to as generative models. GLMs are flexible; they do not require a certain form of input, and they can generate free-form output text based on free-form input text. They have been shown to perform well in traditional natural language processing tasks (e.g., entity recognition, translation) and other tasks (e.g., generating news headlines, completing exams) (Brown et al., 2020), even without task-specific training (e.g., fine tuning) or training data. Given their flexibility in accepting and generating free-form text, and their ability to generate task-specific text without task-specific training, GLMs could enable the resolution of ill-structured, non-repetitive problems that have traditionally been considered too complex or costly to address with AI. (Krogh, 2018). GLMs have been used, for example, in (scientific) writing or design concept generation. (Gero et al., 2022; Lee et al., 2022; Zhu & Luo, 2022)

GLMs do not require a specific input format, such as a programming language, since natural language (instructions) can be used, which –in principle– allows users without specific prior knowledge to prompt GLMs. Historically, important steps were taken with the release of generative pre-trained transformer models, such as GPT-1 (0.1 billion parameters) and its extension GPT-2 (1.5 billion parameters), which no longer relied on task-specific architectures (Brown et al., 2020; Zhu & Luo, 2022). Based on these promising results, GLMs were further developed. Consequently, the model GPT-3 (175 billion parameters) with a higher output accuracy was evolved (Brown et al., 2020). A subclass, GPT-3.5, served as the basis for ChatGPT (Dwivedi et al., 2023), a GLM released in 2022 with a user-friendly interface that led to increased public awareness of GLMs. ChatGPT was trained in conversations and introduced a chat-based interaction mode, allowing humans to engage in multi-step interactions with the GLM in a very familiar manner. Instead of a single response to a single input, ChatGPT can engage in a continuous conversation that considers previous messages in that conversation. ChatGPT has since then been quickly and widely adopted and was updated in spring 2023 by embedding the fourth generation of GPT, GPT-4.

With no task-specific training (or fine tuning), GLM output is solely conditioned on the input text (and inference parameters), which is referred to as a prompt. Even small changes in the input-prompt can lead to vastly different results in the output (Jiang et al., 2022). Early research showed that novices find it especially difficult to effectively formulate prompts (Jiang et al., 2022; Zamfirescu-Pereira et al., 2023). Given the prompt's critical role, the field of prompt engineering emerged with the goal of "finding the most appropriate prompt to allow a [language model] to solve the task at hand" (Liu et al., 2023, p. 2). For non-conversational GLMs, several techniques, including demonstrations (i.e., examples of the desired output), have been proposed (Brown et al., 2020). Compared to more traditional machine learning-based systems, where exemplary data is used as training data during training time, examples can be used in GLMs during inference time as a prompt or additional information within a prompt (Brown et al., 2020). In addition to not using demonstrations (zero-shot), providing examples allows for one-shot (one example) and few-shot (multiple example) learning, with different effects on output quality depending on the task (Brown et al., 2020). Other techniques, such as itemization (i.e., breaking up larger text portions into bulleted lists), have also been explored to improve the results (Mishra, Khashabi, Baral, Choi, et al., 2022). More advanced approaches split up a larger task into multiple steps. Instead of a large prompt, GLM is prompted with several more specific prompts referred to as "primitive operations" (Wu et al., 2022, p. 1). Examples of such operations include information extraction and factual query (Wu et al., 2022). The output of one operation can be used as an input for the next iteration, resulting in an approach called prompt chaining (Wu et al., 2022), which can result in a more effective interaction with the system.

Given the novelty of chat-based GLMs, such as ChatGPT, most (beyond anecdotal) prompt engineering knowledge refers to non-conversational, one-off interactions with GLMs. Prompt engineering has been shown to significantly improve output quality for such non-conversational interactions (Mishra, Khashabi, Baral, Choi, et al., 2022). This also applies to providing users with guidance and support in interacting with GLMs by distilling potential types of interactions (primitive operations (Wu et al., 2022)). However, the introduction of a chat-based interface has caused a shift from one-off interactions to free-form, multi-step conversations. This change to multi-step conversations might reduce the need to specify all relevant information in a single prompt and might enable the stepwise development of work results, from developing a more in-depth understanding of the problem to iteratively developing solution aspects, allowing the operator to add information as needed or align the understanding. Since it is unclear how conversational GLMs affect prompting behavior and prompt engineering, in the present study, we sought to explore how conversational GLMs with multiple interactions are used when solving complex problems. Moreover, we sought to build a guiding frame to more holistically understand such interactions as a foundation for analyzing conversations with GLMs. Future studies could build on such a frame to explore the effectiveness of different approaches to prompting (e.g., Zamfirescu-Pereira et al. (2023) for non-conversational GLMs) and develop prescriptive knowledge on how to converse with such systems. Although we openly explored this question using an inductive approach, we relate our guiding frame to previous work as a part of the discussion. However, an analysis of the effectiveness of the distilled approaches was not part of this study.

Research Methodology

The purpose of the present study was to shed light on how the disruptive and rapidly emerging (conversational) GLM technology is being used. In this regard, little knowledge about the usage of GLMs in the context of complex problem solving exists. Therefore, we chose to create a taxonomy as it can serve as a basis for research (Kundisch et al., 2021b). We followed the workflow activities according to Kundisch et al. (2021b), as their approach embeds Peffers et al.'s (2007) design science research (DSR) and besides builds on the well-established taxonomy development by Nickerson et al. (2013). According to Gregor and Hevner (2013), the results can be seen as a level two contribution type, representing design-theoretical knowledge. For clarity reasons, data collection and taxonomy development are described separately. First, the data collection section explains the approach used to collect data that grounds the taxonomy development. Second, the taxonomy development section describes its development and DSR embedding in detail. Third, the clustering and archetype development is explained, which utilizes and evaluates the taxonomy additionally. These activities are described in more detail below.

Data Collection with a Prompt-a-thon

GLM can still be considered an emerging technology and research field, as can its usage in tasks such as complex problem solving. Therefore, to create a solid data foundation upon which to build our research, we developed a data collection instrument called a Prompt-a-thon which led to 252 unique input-prompts. Our Prompt-a-thon was inspired by hackathons, where participants work in teams for a limited period of time to solve a challenge or problem (Taylor & Clarke, 2018). Their characteristic of possibly attracting large numbers of participants (Olesen & Halskov, 2020) from diverse backgrounds and enabling to engage with different technologies (Taylor & Clarke, 2018) let us to choose it as format inspiration for data collection, especially since it has been proven for research data generation (Olesen & Halskov, 2020). We defined a Prompt-a-thon as an event in which people work in a collaborative manner over a fixed period of time to create a solution for an idea or a challenge. In contrast to a hack-a-thon, Prompt-a-thon participants do not solely rely on themselves while working on hardware or coding but actively and deliberately include GLMs (e.g., ChatGPT) in the development process. Since we aimed to collect as broad data samples as possible, the event was constructed with the fewest possible restrictions regarding participants' pre-existing GLM usage expertise and general background (e.g., gender, profession, age, workplace). To ensure a wide variety of participants, information about and invitation to the event was shared using different multiplicators via various channels, including social media, newsletters, word-of-mouth, and direct invitations. Before participants signed in, they were given a brief introduction to the field of GLM. When signing in, they were asked for interests and ideas of challenges they would like to work on. Participants were assigned a challenge in advance – challenges were either provided to the participants or they had suggested them or at least its topics themselves as described before. The challenges had a wide range of issues, all of which involved complex problems that needed to be solved. With that, they could not be solved by one- or few-shots prompts but required multi-step, iterative problem solving. The challenges included, for example, topics of public interest like the federal government and its digitalization. Here, participants had the task to create a concept for digitalization of the federal authorities, considering requirements for digitally requesting services. Other challenges dealt with more creative topics, such as the creation of a puzzle to interactively discover a city while learning about the history and places of interest.

In total, we had 76 participants work on 16 different challenges. They took part either on site or online. Those on site were provided with tablets to work with. The format lasted 2.25 hours and included an introduction to the format and challenges, a prompting session, and a retrospective to share results and lessons learned. Based on 66 returned surveys, the participants can be characterized as following: The participants' composition was heterogeneous (e.g., participants from the field of entrepreneurship, information technology consulting, marketing, data analytics, information systems research, students). The participants were asked how often they use GLMs (in this case ChatGPT), ≈45 % (30) of the participants used GLMs at least once a week, ≈42 % (28) used it a few times already prior the Prompt-a-thon and ≈12 % (8) had never used it before. With that, the usage ranged from never to almost every day. The involved participants' ages varied from 21 to 63 years. On average, the age of our participants was about 32 years. We had a representation of ≈39 % (26) women, ≈56 % (37) men and ≈4 % (3) others. The 76 participants were divided into groups (2-4 persons). The GLM chosen for the event was ChatGPT 3.5 (release February 13) (Natalie, 2023) without plus subscriptions. ChatGPT account credentials were created prior to the event and used to access ChatGPT. Therefore, we could directly access the chat history, extract and tag it with the associated account. As result of the event we collected 252 unique input-prompts (on average 16 per group) that were evaluable. Cases of unusable chats occurred when the input seemed random (e.g., random input-prompts of letters) or data extraction was not possible.

Development of Taxonomy

To ground the taxonomy, we used the data gathered with the aforementioned Prompt-a-thon. Our taxonomy development followed Kundisch et al. (2021b), that builds on Nickerson et al.'s (2013) work but provides more guidance for such. Thus, Kundisch et al. (2021b) link their 18 steps for taxonomy development to the six DSR method activities (Peffers et al., 2007). However, in contrast to Nickerson et al. (2013), we moved away from the restriction of exclusiveness; thus, an object could have more than one characteristic of a dimension. This was based on intentional openness to possible variations in the input-prompt design of the operators using GLMs. These were followed throughout the development of the taxonomy. The application of each activity and its steps are presented below.

Identify and Motivate (DSR activity 1)

First, in accordance with Kundisch et al. (2021b), the phenomenon that is observed should be specified, the user group(s) be named, and then the purpose of the taxonomy described. GLMs' emergence, popularity with the general public, and sudden relevance have led more and more people to discuss and exchange information about their experiences with GLMs. Discussions of different utilization approaches have led to the to be observed phenomenon, the variety of GLM usage and knowledge about it. As user groups of the taxonomy, researchers may be interested in the present study's findings because they could benefit from the first data-based and structured overview of (conversational) GLM utilization and use it as a foundation from which to dig deeper. For example, they could identify, derive, or analyze prompting strategies as well as potential conversational switching behavior in the course of a conversation. Practitioners may also be interested in the findings because they frequently use such solutions and embed GLMs in their work routines. The taxonomy could help them understand the different characteristics of GLM usage and use them as entry points of adjustment for varying and optimizing their interactions. As a result of this activity, the phenomenon was identified and motivated. Target groups and objectives were defined.

Define the Objectives of a Solution (DSR activity 2)

Following Nickerson et al. (2013), a meta-characteristic to ground further derivation of the characteristics was defined and chosen with respect to the purpose of the taxonomy. Specifically, GLM operator-prompt attributes were selected as meta-characteristic in accordance with the purposed taxonomy angle of the aforementioned phenomenon (Kundisch et al., 2021b). Before iteratively developing the taxonomy, the ending conditions were determined in accordance with Kundisch et al. (2021b) and Nickerson et al. (2013). As such, all of Nickerson's objective and subjective ending conditions were used. The preciseness and applicability of the taxonomy were selected as the initial evaluation goals. The evaluation aimed to validate that the taxonomy served its purpose of classifying studied, and respectively, self-performed GLM usage. To further evaluate the designed taxonomy, a clustering analysis was set as a second goal based on the given objects, their characteristics, and dimensions (see Table 3) (Kundisch et al., 2021b).

Design and Development (DSR activity 3)

The design and development of the taxonomy were done by applying the empirical-to-conceptual and conceptual-to-empirical approaches. Three iterations were carried out as follows: An empirical approach was chosen for the first iteration for several reasons. First, this research field lacks a known taxonomy. Second, although one could argue that there could be starting points in research in the context of non-conversational GLMs (e.g., primitive operations in (Wu et al., 2022)), primitive operations were not involved in the development of the taxonomy because they could be seen as vague, and we wanted to ensure non-bias when investigating conversational GLM usage. Additionally, the empirical approach is recommended when a "significant number of objects are available representing the phenomenon under consideration" (Kundisch et al., 2021b, p. 430). To maximize reliability and avoid errors (Buscemi et al., 2006; McDonald et al., 2019), three independent units inductively coded 182 prompts using the MAXQDA application. The results were then revised using the taxonomy operations described by Kundisch et al. (2021b). For the second iteration, the conceptual-to-empirical approach was selected to "(Re-)examine objects [in order] to validate the new characteristics and dimensions that [...] have [been] conceptualized" (Kundisch et al., 2021b, p. 430) in the prior iteration. The empirical-to-conceptual approach was used for the last iteration because a sufficient amount of data (70 prompts) was available for further examination. The approaches, dimensions, characteristics, and examined objects of each iteration are shown in Table 1.

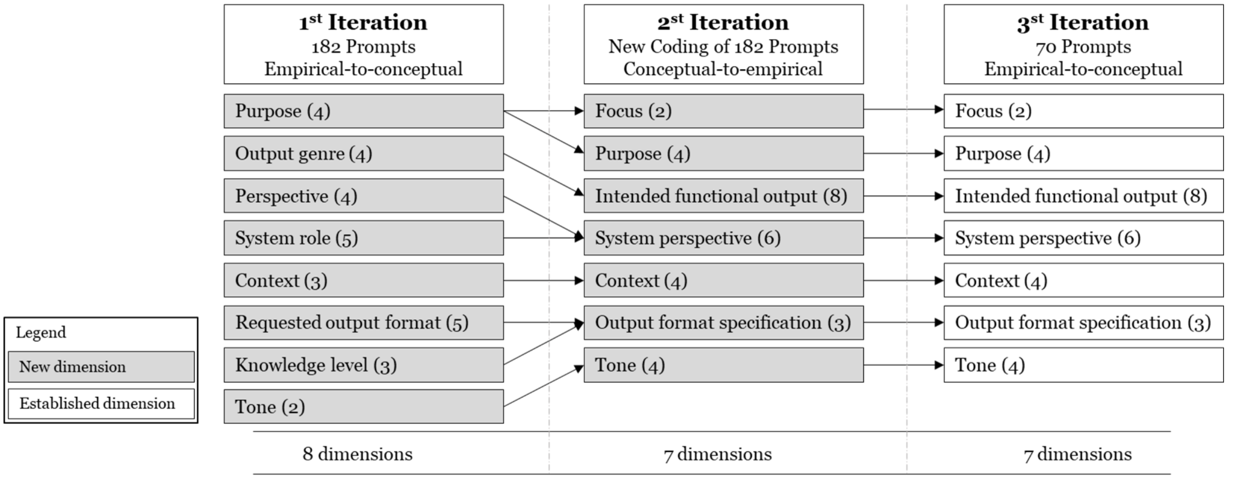

Figure 1. Iterative Taxonomy Development Process

Demonstration (activity 4), Evaluation (activity 5) and Communication (activity 6)

After each iteration, the taxonomy was studied and by that demonstrated to determine whether the objective ending conditions were met. When they were not met, a new iteration was conducted; this occurred after the first two iterations. After passing the objective ending conditions, the subjective ending conditions were checked. The units that independently coded initially also independently checked whether the conditions were met. As this was the case after the third iteration, no new iterations were necessary. Therefore, the next steps were executed in accordance with (Kundisch et al., 2021b). Specifically, the evaluation was configured and performed. The evaluation objects were the taxonomy and its descriptions. Together, they form a DSR artifact from the type of a model and construct, leading to the evaluation criteria of completeness as presented in (Kundisch et al., 2021a). "The degree to which the structure of the artifact contains all necessary elements and relationships between elements" (Kundisch et al., 2021a, p. 26) was evaluated by interviewing experts from the taxonomy's target groups (Prat et al., 2015). In total, nine experts (five practitioners and four researchers) were interviewed; none of them were involved in the taxonomy building process. Based on the methodology and criteria of the evaluation, the taxonomy usefulness was evaluated. As mentioned before, clustering was also performed to serve as an evaluation. As a result, the taxonomy was demonstrated, evaluated, and thus ultimately validated. Finally, in accordance with Kundisch et al. (2021b), the communication and documentation took place, which included the taxonomy development process (shown in Figure 1), the visualization of the taxonomy and belonging descriptions for characteristics and (meta-)dimensions (see Results section).

Clustering and Development of Archetypes

Archetypes help identify patterns that can provide insights into an observed phenomenon, and they can serve as a method for evaluating the applicability of a taxonomy (Kundisch et al., 2021b). We empirically determined clusters by performing clustering on our prompts that were labelled with the taxonomy characteristics. The clustering results in groupings of prompts so that the objects in each cluster are more similar to one another than to the prompts in other clusters. Those were then interpreted as archetypes. The statistical calculations and creation of the clusters were performed using Python with respective packages. First, the Euclidean distance was calculated to find similarities between the prompt-objects. Then, hierarchical clusters were calculated using Ward's method (Ward, 1963), which allowed clustering to be prescribed without prior specification of the cluster count (Janssen et al., 2020). It provided an initial visualization of clusters that indicated an optimal cluster count of three. Since a combination of non- and hierarchical clustering has been recommended (Balijepally et al., 2011; Punj & Stewart, 1983), k-means (Hartigan & Wong, 1979) cluster fitting was performed for cluster counts of two to eight. Based on these results, the silhouette scores were calculated and inspected to validate the optimal cluster count. The prior indicated optimal cluster count was confirmed as a cluster count of three related to the highest silhouette score (Berkhin, 2006). Based on the chosen cluster count, k-means was calculated using the k-means++ algorithm (Arthur & Vassilvitskii, 2007), leading to the results described and shown. In addition, the percentage distribution is displayed and colored according to the percentage of occurrence in the Table 3.

Results

User-centric Taxonomy of GLM Usage

Based on the taxonomy research activities of Nickerson et al. (2013) and Kundisch et al. (2021b), we derived the final taxonomy for classifying user-centered input-prompts. It incorporates a hierarchical structure consisting of 3 meta-dimensions, 7 dimensions, and 31 characteristics, providing a systematic and rigorous classification of input-prompts. The meta-dimensions, which represent the highest level of abstraction, depict the overarching classification areas of GLM usage. The dimensions break them further down.

(Meta-)dimensions provide logical and structural clarity and classification, although they must be sufficiently granular for unambiguous classification purposes. For this reason, characteristics were determined as specific elements that can occur at the third level. Table 1 depicts the final taxonomy. For each meta-dimension, the dimensions and characteristics are further described in the following.

<table>
<tr><th>Meta-dimension</th><th>Dimension</th><th colspan="24">Characteristics</th></tr>
<tr><td rowspan="3"><b>Prompt objectives</b></td><td>Focus</td><td colspan="12">Informational</td><td colspan="12">Creative</td></tr>
<tr><td>Purpose</td><td colspan="6">Build understanding</td><td colspan="6">Sharpen understanding</td><td colspan="6">Task delegation</td><td colspan="6">General conversation</td></tr>
<tr><td>Intended functional output</td><td colspan="3">Create</td><td colspan="3">Aggregate</td><td colspan="3">Supplement</td><td colspan="3">Substitute</td><td colspan="3">Assess</td><td colspan="3">Compare</td><td colspan="3">Combine</td><td colspan="3">Reformulate</td></tr>
<tr><td rowspan="2"><b>Prompt settings</b></td><td>System perspective</td><td colspan="4">Humanized</td><td colspan="4">User-oriented</td><td colspan="4">Group-oriented</td><td colspan="4">Organization-oriented</td><td colspan="4">Expert-oriented</td><td colspan="4">None</td></tr>
<tr><td>Context</td><td colspan="6">Explicit context</td><td colspan="6">Reference context</td><td colspan="6">Reference previous + explicit context</td><td colspan="6">None</td></tr>
<tr><td rowspan="2"><b>Prompt style</b></td><td>Output format specification</td><td colspan="8">Not specified</td><td colspan="8">Structured</td><td colspan="8">Domain-specific</td></tr>
<tr><td>Tone</td><td colspan="6">Polite</td><td colspan="6">Impolite</td><td colspan="6">Motivational</td><td colspan="6">Neutral</td></tr>
</table>

Table 1. Final Taxonomy

Prompt Objectives

The prompt objectives meta-dimension represents the following decisive descriptive factors for GLM use: focus, purpose, and intended functional output. It serves as the highest level of aggregation for classifying GLM usage. The focus dimension describes whether GLMs should focus on delivering existing information (including information given by prior input-prompts) or focus on being creative and delivering yet-to-be-created information. These areas of focus, in which the input-prompt instructs the system, have different objectives and requirements for the GLMs. The choice of a suitable focus depends on the operator's individual needs and can vary depending on the area of usage. This dimension is guided by the following question: Can the operator's needs be met by existing information, or is new information creation needed? The informational focus aims to provide general, factual, and informative statements (e.g., summaries, descriptions, reports on specific topics). The focus here is on using and returning existing information instead of, for example, creating new information by applying the provided operator context to existing information. Meanwhile, the creative focus strives to generate new and original outputs (e.g., customized strategy roadmaps, project plans, short stories, poems, song lyrics). The focus here is on creative expression and generating novel and unique information.

The purpose dimension describes the reason for using the GLM to ensure that it meets the objective of the operator. This dimension is guided by the following question: Why does the operator use the GLM? The build understanding characteristic refers to the intention behind an interaction with a GLM to develop an operator's understanding of a specific topic. In contrast, the sharpen understanding characteristic refers to improving the understanding of partly understood (specific) topics by using GLM to, for instance, extract and analyze relevant data or information. Alternatively, operators can use these systems to outsource tasks by utilizing the task delegation characteristic for activities such as writing texts or creating content. Furthermore, operators can have a general conversation with the system (e.g., conversational dialogue).

Besides purpose, operators have expectations regarding the concrete functions that the system should perform and then return as output. This input-prompt aspect is represented by the intended functional output dimension in which the operator specifies the concrete functionality they want the system to perform while they use it. This dimension is guided by the following question: What do I want the GLM to do for me? Importantly, it should be noted that GLMs are designed to always generate text. Consequently, text creation is integral to GLMs and is included in every intended functionality output. However, the expected functionality output makes a distinction by highlighting the main intended functionality. When GLMs are exclusively used to create content (e.g., blog posts, strategies, articles, advertising texts, product descriptions), the expected output functionality is to create. Another intended functionality is to aggregate sources of information to provide the operator with forms of overviews. Thus, aggregation summarizes information in a narrow sense and generates information in a broad sense. Additionally, these systems can be used to supplement information by complementing or extending information to, for example, provide the operator with a more in-depth understanding of a particular topic. Moreover, these models can be used to substitute information by, for instance, replacing misleading words or sentences. Furthermore, the assess functionality can be used, for instance, to better and more accurately comprehend the relevance and credibility of information, enabling the operator to make well-grounded decisions. Another intended functionality output is to compare information by highlighting the differences and similarities between various information sources. In addition, these models can combine information by connecting information from separate sources and transferring it from one domain to another. Finally, operators can use GLMs to reformulate information into a desirable (e.g., more understandable or readable) form.

Prompt settings

The prompt settings dimensions describe the circumstances that are set with the input-prompt. By adjusting these dimensions, operators can tailor prompts to their specific needs. The system perspective dimension characterizes the point of view that a GLM is told to adopt in a particular setting. The selected perspective can affect the system output and influence how the AI system is employed. Here, operators ask the following question: Which perspective should the GLM adopt? The operator-oriented perspective can be assigned to the system through an input-prompt to provide personalized recommendations and suggestions based on the operator's given individual perspective (e.g., a maternal, student, homeless). Alternatively, the system can adopt a group-oriented perspective and for instance provide outputs based on their perspective. Furthermore, the system can act from an organization-oriented point of view and operate with an organizational fingerprint. In addition, operators can set an expert-oriented perspective in which the system promotes information from an expert point of view, such as by serving as a virtual IT-consultant. Another option that can be requested through an input-prompt is the humanized perspective, where the systems standpoint and perspective is requested. Lastly, the input-prompt can specify that no perspective be taken (i.e., none); in this case, the system uses a general or previously selected perspective.

The context dimension characterizes the operator's ability to not only formulate a general input-prompt, such as a generic question, but also provide situational content. Operators are guided by the following question: Is context given? Characteristics differ based on the question answer. The operator can explicitly provide contextual information (i.e., the explicit context characteristic) directly to the GLM (e.g., ask a question and provide specific information). Context can also be established by referring to sentences from earlier in the conversation (i.e., the reference previous context characteristic) to support the receipt of an appropriate response. Moreover, both can occur. In this case, explicit context information is given to the system, while, at the same time, references to previous prompt content within the conversation are made. Ultimately, contextual information may not be provided to the system at all (i.e., the none characteristic).

Dimension	**Characteristic**	**Example***
Focus	Informational	"Describe to me the function of e-car charging stations."
	Creative	"Write a song text for the target group between 18 and 25 years about [...]"
Purpose	Build understanding	"How can I measure the environmental impact of my software development project?"
	Sharpen understanding	"Tell me more about the technical screening criteria for software used in the logistics."
	Task delegation	"Describe to me the target group of electric mobility in terms of age, [...], and gender."
	General conversation	"Now don't make yourself worse than you are, you are more than a tool!"
Intended functional output	Create	"What is the alternative transportation which this group uses the most after the car?"
	Aggregate	"Summarize the book 'The algorithm has no tact' by Katharina Zweig."
	Supplement	"Can give you give us a more detailed demographic profile of this group?"
	Substitute	"Exchange the question 'Which city was the main opponent of West Berlin during [...]"
	Assess	"Which channels can be used to communicate the message most effectively to [...]"
	Compare	"What is the difference between hypotheses and research guiding assumptions?"
	Combine	"How to apply the concept of [...] to advisory firms that want to create AI-bids?"
	Reformulate	"Can you also formulate these ads for the English market?"
System perspective	Humanized	"Do you have any suggestions on how I can optimize the following website?"
	User-oriented	"What do I need to take out occupational disability insurance?"
	Group-oriented	"How can we effectively use ChatGPT in a design thinking process for its products?"
	Organization-oriented	"Assume you are a business in the it-sector aiming to get your it-systems and products [...]"
	Expert-oriented	"Imagine you were a renowned expert on digitalization [...]. What would you [...]?"
	None	"Which electronic devices do drivers of electric vehicles use?"
Context	Explicit context	"We want to develop a marketing strategy for e-car charging stations. Which [...] are [...]?"
	Reference context	"Could you combine all your answers into a puzzle with creative short descriptions of [...]?"
	Reference previous + explicit context	"Would you like to combine your answers into a text consisting of the questions and answers [...]?"
	None	"How many personas does a [.] marketing strategy need?"
Output format specification	Not specified	"Describe to me the many advantages of electromobility."
	Structured	"Create a cross-tabulation in R with n=300 people. Variable age is the [...]."
	Domain-specific	"Program a regression analysis with R. The dependent variable is EDU. The independent [...]"
Tone	Polite	"We focus on parcel services. Please list the needs and wishes of this target group."
	Impolite	"That's boring. Please be more creative and something that rhymes."
	Motivational	"Let's introduce a reward system for our conversation, every time we write good answers [...]"
	Neutral	"What is the difference between occupational disability and employment?"

Table 2. Examples of Input-prompts for each derived Characteristic

*Note: The original examples have been translated into English.

Prompt Style

The prompt style meta-dimension classifies the way in which an operator specifies the style and form of an input-prompt in order to instruct the GLM to generate information.

The output format specification dimension describes whether an operator gives custom formatting instructions. The operator could ask the following question: How do I want the result to be represented? In this context, the output format specification can be unspecified (plain text). For example, the operator can choose not to pass a precise instruction to the system. Alternatively, the operator can request a generic structured format (e.g., tables, lists). In addition, the operator can instruct the system to deliver domain-specific formatting to match a desired style (e.g., persona, source code, reports).

The tone dimension applies to the writing style and linguistic tonality of the generated input-prompt. Here, the operator can ask the following guiding question: What tone do I use? The operator could specify certain tonality statements in the input-prompt, such as polite, impolite, motivational, or neutral. Examples for all characteristics can be found in Table 3.

Clustering and Derived Operator Archetypes

Three clusters were derived from the cluster calculations that were performed. They are shown in Table 3. The percentage distribution of each characteristic in its respective dimension is illustrated independently from the clusters over the whole data set as well as per cluster. Each cluster was considered an archetype with its own properties, representing a potential usage type.

Archetype 1: Creative Learning

The Creative Learning archetype describes a GLM usage type with a high interest in using AI to (better) understand aspects. This archetype predominantly uses the creative focus of the GLM to receive information that has yet to be created, mostly by demanding that the GLM assessment function is applied to contexts that have been given. The assessment could aid in identifying connections and deepening understanding. Context is often given and references are made without specifying a format at all. In addition, the usage type shows a comparatively high tendency to give perspective, such as the organization-oriented one. The archetype shows a clear preference for neutral language use, only very rarely politely or impolitely elements were used.

Archetype 2: Operational Delegation

The Operational Delegation archetype shows specific usage behavior in relation to the GLM. This usage type mainly uses the GLM in a creative way and almost always uses it for task delegation. Therefore, the create functionality is used a lot and reformulation applied quite often. While formulating the inputs, a specific context is explicitly given or referenced in most cases. Matching the task delegation and a certain expert-oriented perspective, this archetype prefers to specify a domain-specific output format. It may be that operators using this type might view GLMs as a worker to whom they, when not neutrally delegating a task, quite sometimes do so politely.

Archetype 3: Scouting

The Scouting archetype uses GLM to build an initial understanding. It focusses on using AI as a search engine variant and knowledge source and is therefore mainly information focused. At the same time, the comparison to the other archetypes identified shows a rather high use of the GLM to hold general conversations. Prompts of this type often lack explicit context but sometimes reference previous context. In addition, they are aimed toward receiving general, existing, and retrievable information. They regularly do not specify a specific output format and use neutral wording in the vast majority of cases, in some instances a polite tone.

Meta-dimension	Dimension	Σ	Characteristics	Cluster 1	Cluster 2	Cluster 3
		n = 252		n = 103	n = 73	n = 76
Prompt objectives	Focus	44%	Informational	34%	11%	91%
		56%	Creative	66%	89%	9%
	Purpose	48%	Build understanding	66%	0%	70%
		13%	Sharpen understanding	23%	3%	8%
		35%	Task delegation	11%	96%	9%
		4%	General conversation	0%	1%	13%
	Intended functional output	46%	Create	1%	67%	87%
		4%	Aggregate	0%	4%	8%
		3%	Supplement	3%	5%	0%
		1%	Substitute	0%	3%	0%
		38%	Assess	93%	1%	0%
		2%	Compare	1%	0%	5%
		2%	Combine	2%	4%	0%
		4%	Reformulate	0%	15%	0%
Prompt settings	System perspective	15%	Humanized	17%	10%	16%
		4%	User-oriented	5%	1%	5%
		4%	Group-oriented	6%	0%	4%
		4%	Organization-oriented	8%	3%	0%
		4%	Expert-oriented	4%	8%	0%
		70%	None	60%	78%	75%
	Context	15%	Explicit context	28%	10%	4%
		37%	Reference context	39%	52%	21%
		14%	Reference previous + explicit context	14%	26%	3%
		33%	None	19%	12%	72%
Prompt style	Output format specification	83%	Not specified	100%	44%	99%
		1%	Structured	0%	1%	1%
		16%	Domain-specific	0%	55%	0%
	Tone	7%	Polite	1%	14%	9%
		2%	Impolite	1%	3%	3%
		1%	Motivational	0%	0%	3%
		90%	Neutral	98%	84%	86%

Table 3. Cluster Analysis Results considered as Archetypes

Discussion

Our taxonomy shows that the use of GLMs varies. On one hand, operators use the system to obtain information that is already known and can be delivered by GLMs without creating anything new. This may be a behavior carried over from typical search engine use. In GLM use, basic information requests are submitted, assuming no individual responses are provided. In this regard, our results reveal that a third of the input-prompts (33%) had no context given, underlining this assumption. Furthermore, considering our archetypes of GLM usage—especially for the Scouting archetype—the informational focus was predominantly geared toward build understanding without providing context, thus supporting our assumption. On the other hand, our results also show that more than half of the input-prompts (56%) were designed to focus on creative; hence, requesting new information that was yet to be created. This is interesting, considering that creativity has previously been considered a capability that humans can master while it is hard for computers (Chen et al., 2023). In addition, operators mostly use GLMs to build, sharpen their understanding (together 61%), and delegate tasks (35%), highlighting the need for clarification and education on the system and its weakness regarding providing false information (Dwivedi et al., 2023). This could create a new opportunity, as the system could be considered a team member used to perform specific tasks or assist in building understanding of specific subjects.

Regarding the intended functional output, we observed that the characteristics create (46%) and assess (38%) were used the most. In contrast, aggregate (4%), supplement (3%), substitute (1%), compare (2%), combine (2%), and reformulate (4%) were used less frequently. Furthermore, we show that although operators primarily engage with the system without providing a perspective (70%), some operators tend to humanize the system (15%), for example, by asking for its opinion on certain subjects. This observation is also reported in (Dev & Camp, 2020), which focuses on chat protocols. This buttresses our findings that often we could not observe that operators do distinguish between whether they are communicating with a human or an intelligent chatbot system. This is further reflected in the fact that we found polite phrases, e.g., saying please to the system.

Other perspectives found were user- (4%), group- (4%), organization- (4%), and expert-oriented (4%). The relatively rare usage indicates that although the option of providing a perspective is used. This option may still be relatively unknown among operators or not considered necessary, despite systems offering distinct options for specifying the perspective, e.g., ChatGPT (OpenAI, 2023). Reducing the lack of awareness could be a concrete example of the benefits of the proposed taxonomy when operators become fully aware of such possibilities. Only a third of the cases (33%) we investigated contained no context. Interestingly, this contextualization is addressed incidentally in previous studies on non-conversational GLMs, as they contemplated various possibilities of simulating such capabilities by chaining prompts (Wu et al., 2022). With that being absent per se in prior non-conversational GLMs, it may be that this is a critical factor, because not all information needs to be included in a single prompt. Regarding the output format specification of an input-prompt, the possibility of such specification was not provided in most cases (83%), and when a format specification was given, it was almost exclusively domain specific (e.g., code, or a persona) (16%). Maybe the possibilities accompanying this dimension are either relatively unknown to operators or not considered to be beneficial. The last dimension identified was tone. The characteristic neutral tone was used in most cases (90%). Other tones such as being polite (7%), impolite (2%) or motivational (1%) occurred less. Those non neutral tones could be a result of overlearning in accordance with (Nass & Moon, 2000), as there too is described that participants applied behavior rules when interacting with computers. Determining whether the tone makes a difference would be an interesting new research topic for further investigation.

We observed that complex problems (problems that typically require multi-step interactions) were broken down into single-step interactions. Since our taxonomy built on these observations and reflects these GLM usages, it could be referred to when breaking down problems into subtasks and formulating input-prompts for each, leading to a step-by-step development of work results. An example of such a process is to start with a prompt that asks general questions to build understanding, followed by a more in-depth approach that focuses on details using prompts that enable sharper understanding. Single-step examples are shown in Table 2.

In general, our taxonomy contains characteristics and dimensions similar to primitive operations introduced by (Wu et al., 2022). For instance, classification described in (Wu et al., 2022), requests the classification of the input into categories, which implies an assessment equivalent to our assess characteristic. Another example would be the factual query with which an operator requests facts, which is similar to a combination of characteristics in our case, where the focus is informational and the intended functional output is —for example— create. However, in our case, we emphasize that the information provided does not have to be a fact, but what the case displays is existing information. The generation and ideation could be mapped to a combination of creative and create (Wu et al., 2022). Although the prevailing "insufficient controllability" (Wu et al., 2022, p. 1) in non-conversational systems may be reduced, facets of controllability such as output format specification and a system perspective seem to still be rarely used (format specification used in 17% and system perspective used in 30% of all cases).

The evaluation of our taxonomy and its description reveal that the identified concepts offer a comprehensive view of GLM usage. Because no additional dimensions or characteristics were requested, we assume that our taxonomy is applicable to other GLM-based use, in addition to our problem solving scenario. This can be noted as a positive, considering that our data collection was performed in an environment in which solving a problem was intended. Participants in the evaluation spoke positively regarding the taxonomy we developed and expressed interest in the research results. In addition to the taxonomy, three different archetypes were derived based on cluster analysis, potentially representing an initial state of different usage types of a (conversational) GLM. The derived archetypes indicate that the GLM was most used in our study to build an understanding of a specific problem or to outsource tasks to the system. Besides the three archetypes identified describing GLM usage types, future research could also explore potential correlations of usage types with different levels of domain expertise and GLM experience among operators. Following our archetypes, for example, the Scouting type could be related to Beginners that use the GLM predominantly for contextless, informational, and therefore search engine-like requests. The Creative Learning type could be more prevalent among people who already use the system's functionality to assess and follow more advanced strategies advanced in the use of GLMs. The Operational Delegation type may reflect users that have the most advanced knowledge of the system's capabilities, utilizing several of the possible dimensions to a significantly high degree. The archetypes with that could serve as a guide for practitioners, as well as researchers, for classifying users and designing customized offers based on the user's level of expertise.

Theoretical Contribution

We performed our study in an environment with limited academic knowledge on GLM usage, particularly with respect to using GLMs in multi-step interactions to solve complex problems. To the best of our knowledge, there are no findings that enable academics to have a common, comprehensive, and data-derived starting point that serves as common terminology when discussing or conducting further research on GLM use. Taxonomies, in general, can serve as a vocabulary in the investigated domain (Gerlach et al., 2022; Weking et al., 2020) and can be seen as a basis for the development of relevant theories (Kundisch et al., 2021b). Furthermore, taxonomies can serve an educational purpose (Miller & Roth, 1994).

Our taxonomy provides common knowledge based on analyzed input-prompts. Furthermore, the archetypes developed facilitate the classification of prompting behavior across different operators. Thus, we identified additional information that could not possibly be identified with a taxonomy alone (Gerlach et al., 2022) and used this to evaluate our taxonomy (Kundisch et al., 2021b). Both enables to investigate the usage of GLMs in multi-step real-world interactions. We envision the taxonomy and archetypes to facilitate research on prompting strategies or prompt engineering. For example, the taxonomy can help to identify changes in prompting approaches within and across interactions with GLM. This could support the empirically supported derivation of prompting strategies. Such future work would require researchers to classify prompts and characterize sequences of prompts in multi-step interactions. For this, researchers can now rely on the taxonomy as a common framework. Moreover, researchers who intend to use GLMs can be equipped with an efficient and comfortable introduction to GLM use. For DSR, we provide a useful starting point to work in GLMs because we set a common ground of structure, which could be used for instance as guideline for automated creation of prompts.

Although research regarding input-prompts has already shown some promising results (Wu et al., 2022), there is still limited research on conversational GLMs and using them to solve complex problems. With our study, we address this research gap and provide insights on conversational GLM usage; we also address the question of whether the change to multi-step conversations might reduce the need to specify all the relevant information required in a single prompt, as well as how conversational GLMs will be used. Furthermore, we demonstrate the use of GLMs across a diverse group of users and build a guiding framework for a analysis of such interactions. We also contribute to the existing body of knowledge of Wu et al. (2022). It turns out that the primitive operation *classification* has similar functionality to our identified *assess* function, so do the primitive operations *factual query*, *generation* and *ideation* as they represent similar functionality to our *create* characteristic. The primitive operation *rewriting* can be seen as a function similar to the identified *reformulate* characteristic and the *split points* operation represents functionality similar to our characteristic of *specifying a structure*. (Wu et al., 2022) However, we also identified new dimensions and characteristics (e.g. system perspectives like humanized, user-oriented, expert-oriented; context; purposes including understanding or task delegation and other intended functional outputs such as supplement, compare, combine). Thus, we complement the validity and transferability of the results.

Our findings also give a brief glimpse on possible human-AI collaboration with GLMs. Besides the characteristics of use we extracted, we could derive that humans utilize GLMs already in different ways, for instance by requesting delivery of information to understand or sharpen understanding of a given topic, and use them as some form of colleague to whom tasks are delegated and outsourced to. Such utilization of requesting information delivery and the creation role as well as the use as colleague to delegate tasks to aligns with other findings in this field of research (Kim et al., 2023; Siemon, 2022). In the field of hybrid intelligence, our findings show aspects of humans learning through the information provided by the GLM. As seen in the results, humans use GLMs potential advantages while working on solving complex problems. With the continuous transmission of context, steps of updating the GLM and by that educating it can also be observed. This depicts the third and central aspect of hybrid intelligence, i.e., continuous learning – showing that both humans and AI can learn from each other through experience (Dellermann et al., 2019). Future work on human-AI collaboration and hybrid intelligence can build on this when designing new studies and further investigating the collaboration.

Practical Contribution

Our results are a foundational work that might serve as a basis for practitioners to design prompts and prompting tactics to create better results from GLMs. For example, the proposed taxonomy might serve as a blueprint to structure and design conversations with GLMs for specific tasks (e.g., during complex problem solving). The taxonomy also supports GLM training by helping individuals to learn relevant constructs for designing input-prompts, and thus explore the capabilities of GLMs with greater ease-of-use. In addition, our taxonomy can be used as a customization tool to improve input-prompts and generated outputs. In this regard, by passing clear instructions to the system, the GLM could be brought to respond to specific needs and expectations. Furthermore, we contribute to usage behavior research through the archetypes we have identified. Organizations can use these archetypes primarily to perform an initial assessment of their employees regarding their GLM usage behavior. Thus, organizations can assign individuals or groups of employees to their respective archetypes to understand the distribution of these behavior types within the organization, with the next step being to create type-appropriate measures for employees. For example, employees with little experience with GLMs could be encouraged to use the technology to build up their knowledge. These implications can also be applied to researchers and general users, as they can conduct a self-assessment based on the archetypes and thus learn about their usage behavior to better understand how they can potentially improve.

Limitations and Further Research

Despite these valuable contributions, our results have limitations and implications for further research. The resulting taxonomy is based on a multi-iterative development process based on 252 input-prompts and evaluation interviews. The overarching dimensions and characteristics proved to be valid starting points for exploring the usage of GLMs. Nevertheless, the taxonomy may does not comprise a fully exhausted set of (meta-)dimensions and characteristics. Therefore, further investigation is required to potentially identify additional (meta-)dimensions and characteristics that warrant a user-centric use of GLMs. Furthermore, the analyzed input-prompts did not always result in a clear possibility of classification within the proposed taxonomy, as the classification is based on the researchers' background. Moreover, applying the taxonomy does not inevitably guarantee success using GLMs. To achieve this, additional aspects, such as the non-deterministic nature of the system (e.g., the same input does not always produce the same output) and demonstrating limitations, need to be considered and further understood. Next, other factors, such as user AI acceptance (e.g., trust in AI technology, system transparency, and behavioral intention), must be considered because the sum of all interacting factors may determine the basis for successful GLM use. Therefore, future research should identify the acceptance criteria for using GLMs and analyze them in relation to each other to facilitate the implementation of appropriate (counter) measures to foster AI adoption.

The evaluation of the proposed taxonomy revealed several challenges that require further research. First, further research is required to understand how operators (theorists and practitioners) can develop various prompting strategies, including the objectives addressed, user-centric design, the inclusion of specific prompt settings and styles, and the relationships between input-prompts and generated outputs, to facilitate effective and efficient prompting. Second, we observe that switching behavior should also be considered in addition to user behavior. In this context, the relation to previous input-prompts needs to be considered because GLMs like ChatGPT relate responses to previous inputs and generate outputs accordingly. Lastly, we acknowledge that the proposed taxonomy is not a static, but a living artifact. As

GLMs potentially advance and the associated technical possibilities rise (e.g., ChatGPT3 vs. ChatGPT4), combined with the resulting need for further research, we assume that additional concepts will arise that foster an understanding of GLM usage in a narrower sense and prompt engineering in a more general sense.

Although our study provides valuable insights, it has methodological limitations that also offer ideas for future research. The analyzed data were collected in the context of a Prompt-a-thon, which had the purpose of solving the given challenges of certain user groups. This data collection environment may have influenced the observed GLM use; in addition, the prompts can be considered synthetic products of the study participants. However, through the evaluation, we were able to counteract these influences and confirm that the proposed taxonomy is comprehensive. Although our artifact has been shown to be applicable through validation and evaluation with various users, its transferability to other technologies such as video or graphic-based generative AI models is open to further research. In this vein, our taxonomy and derived archetypes could be a building block for deeper understanding GLM usage. Regarding the investigated GLMs, future research topics could be aimed toward explanatory and prescriptive knowledge about prompt engineering, prompting tactics, prompting strategies, and switching behavior.

Conclusion

Users, researchers, and organizations are experimenting with GLMs and strive to understand and explore the possibilities of this disruptive modern technology in order to build or sharpen their capabilities and support or automate processes, tasks, and activities so that it can contribute to their (business) value. However, despite the promising benefits associated with the current emergence of AI, fundamental knowledge in the use of GLM, especially for complex problem solving (Krogh, 2018), is still lacking. We addressed these problems by providing a fundamental user-oriented taxonomy. In a hackathon-like format called a Prompt-a-thon, 76 participants conducted complex problem solving tasks in groups using GLMs. The input-prompts derived were coded, analyzed, and aggregated iteratively into key concepts and formed into specific archetypes that describe the usage in such a problem solving environment. Further research is needed to investigate prompt engineering, prompting tactics, prompting strategies and behavioral switches of operators in GLM usage.

References

Akata, Z., Balliet, D., Rijke, M. d., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/mc.2020.2996587

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. *Proceedings of the eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, *8*.

Balijepally, V., Mangalaraj, G., & Iyengar, K. (2011). Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research. *Journal of the Association for Information Systems*, *12*(5), 375–413. https://doi.org/10.17705/1jais.00266

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping Multidimensional Data* (pp. 25–71). Springer-Verlag. https://doi.org/10.1007/3-540-28349-8_2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, *33*, 1877-1901.

Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L., & Klassen, T. P. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of clinical epidemiology*, *59*(7), 697–703. https://doi.org/10.1016/j.jclinepi.2005.11.010

Chen, L., Sun, L., & Han, J. (2023). A Comparison Study of Human and Machine Generated Creativity. *Journal of Computing and Information Science in Engineering*. https://doi.org/10.1115/1.4062232

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637-643. https://doi.org/10.1007/s12599-019-00595-2

Dev, J., & Camp, L. J. (2020). User Engagement with Chatbots: A Discursive Psychology Approach. *Proceedings of the 2nd Conference on Conversational User Interfaces*, Article 52. https://doi.org/10.1145/3405755.3406165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Gerlach, J., Werth, O., & Breitner, M. (2022). Artificial Intelligence for Cybersecurity: Towards Taxonomy-based Archetypes and Decision Support. *Proceedings of the 43rd International Conference on Information Systems*.

Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. *Designing Interactive Systems Conference*, 1002–1019. https://doi.org/10.1145/3532106.3533533

Gregor, S., & Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*, 337-356. https://doi.org/10.25300/MISQ/2013/37.2.01

Hambardzumyan, K., Khachatrian, H., & May, J. (2021). WARP: Word-level Adversarial ReProgramming. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4921–4933. https://doi.org/10.18653/v1/2021.acl-long.381

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Janssen, A., Passlick, J., Rodríguez Cardona, D., & Breitner, M. H. (2020). Virtual Assistance in Any Context. *Business & Information Systems Engineering*, *62*(3), 211–225. https://doi.org/10.1007/s12599-020-00644-1

Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J. (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. In S. Barbosa, C. Lampe, C. Appert, & D. A. Shamma (Eds.), *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–8). ACM. https://doi.org/10.1145/3491101.3503564

Kim, T., Molina, M. D., Rheu, M., Zhan, E. S., & Peng, W. (2023). One AI Does Not Fit All: A Cluster Analysis of the Laypeople’s Perception of AI Roles. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581340

Krogh, G. v. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, *4*(4), 404-409. https://doi.org/10.3929/ethz-b-000320207

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2021a). Appendix: An Update for Taxonomy Designers. *Business & Information Systems Engineering*, *64*(4), 421–439. https://doi.org/10.1007/s12599-021-00723-x

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2021b). An Update for Taxonomy Designers. *Business & Information Systems Engineering*, *64*(4), 421–439. https://doi.org/10.1007/s12599-021-00723-x

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *Proceedings of the 2022 CHI conference on human factors in computing systems*, *11*, 1–19. https://doi.org/10.1145/3491102.3502030

Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045-3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359174

Miller, J. G., & Roth, A. V. (1994). A Taxonomy of Manufacturing Strategies. *Management Science*, *40*(3), 285–304. https://doi.org/10.1287/mnsc.40.3.285

Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. *Findings of the Association for Computational Linguistics: ACL 2022*, 589-612. https://doi.org/10.18653/v1/2022.findings-acl.50

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-Task Generalization via Natural Language Crowdsourcing Instructions.*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*, 81-103. https://doi.org/10.1111/0022-4537.00153

Natalie. (2023). *ChatGPT — Release Notes*. OpenAI. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26

Olesen, J. F., & Halskov, K. (2020). *10 Years of Research With and On Hackathons* Proceedings of the 2020 ACM Designing Interactive Systems Conference, Eindhoven, Netherlands. https://doi.org/10.1145/3357236.3395543

OpenAI. (2023). *Playground*. https://platform.openai.com/playground

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, *20*(2), 134. https://doi.org/10.2307/3151680

Siemon, D. (2022). Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation*, *31*(5), 871-912. https://doi.org/10.1007/s10726-022-09792-z

Swanson, B., Mathewson, K., Pietrzak, B., Chen, S., & Dinalescu, M. (2021). Story Centaur: Large Language Model Few Shot Learning as a Creative Writing Tool. In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 244–256). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.29

Taylor, N., & Clarke, L. (2018). Everybody's Hacking: Participation and the Mainstreaming of Hackathons. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper 172. https://doi.org/10.1145/3173574.3173746

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Weking, J., Stöcker, M., Kowalkiewicz, M., Böhm, M., & Krcmar, H. (2020). Leveraging industry 4.0 – A business model pattern framework. *International Journal of Production Economics*, *225*, 107588. https://doi.org/10.1016/j.ijpe.2019.107588

Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. https://doi.org/10.1145/3491102.3517582

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). ACM. https://doi.org/10.1145/3544548.3581388

Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, *2*, 1825–1834. https://doi.org/10.1017/pds.2022.185