École Polytechnique Fédérale de Lausanne

Comparing the creative outputs of LLMs

Semester project REACT Group

Maxime Lelièvre

9th June 2024

ABSTRACT

The advent of AI-powered chatbots, particularly large language models (LLMs) like GPT-4, has opened new avenues for enhancing creativity in educational settings. This report investigates the potential of LLMs to assist students during brainstorming sessions, addressing a key pedagogical goal of fostering creativity. While LLMs are designed to predict the next most probable token, which may limit their ability to generate truly novel ideas, this challenge can be mitigated by leveraging their vast general knowledge and occasional hallucinations.

This study sets out to explore how models could be effectively utilized in creative contexts. We developed a methodology to evaluate creativity on Guilford's Alternative Uses Test (AUT), acknowledging its limitations and justifying our methodological choices. The outputs of LLMs were evaluated to understand how different factors, such as varying prompts, influence them. Our findings highlight specific prompting strategies that enhance creativity and those that do not, offering practical recommendations for generating creative ideas on the AUT and mentioning the limitations.

Moreover, we discuss the broader implications of our findings, contemplating the extent to which LLMs could be harnessed to tackle significant, practical and open-ended problems through creative solutions. Based on our discoveries, it is reasonable to speculate that the principles uncovered in this study could be applied to other open-ended problems, paving the way for LLMs to contribute meaningfully to solving real-world challenges. This report not only sheds light on the creative potential of LLMs but also opens doors to future research in leveraging AI for impactful and innovative problem-solving.

All the code and alternative uses generated in this project are available in a public GitHub repository.[1]

[1] https://github.com/Maximelel/LLM_creativity_aut

Contents

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

In recent years, AI-powered chatbots, particularly large language models (LLMs) such as GPT models, have revolutionized various domains, including education. These models are not only capable of providing information and answering queries but also have the potential to foster creativity among students. By offering diverse perspectives and generating a multitude of ideas, LLMs can significantly widen the ideation space for students tackling complex problems. This capability is particularly valuable in educational settings, where encouraging creative thinking is essential for developing problem-solving skills and innovation.

Generating creative ideas typically involves two phases: a divergent phase and a convergent phase. During the divergent phase, individuals are encouraged to generate a wide range of ideas without immediate judgment or criticism. This phase focuses on quantity and variety, promoting free-flowing and non-linear thinking. Conversely, the convergent phase involves refining, evaluating, and selecting the most promising ideas from the divergent phase. This phase emphasizes critical thinking, analysis, and decision-making to identify the most viable and effective solutions. Studies by Guilford 1950 and Runco and Acar 2012 highlight the importance of both phases in the creative process. Our approach in this study ultimately aims to combine both phases where students could leverage LLMs to widen their ideation space (divergent phase) and evaluate creative ideas, which could help them learn what makes an idea creative and support them in their decision-making (convergent phase).

The Alternative Uses Test (AUT), developed by J.P. Guilford, is a widely used measure of creative thinking. It challenges participants to think of as many uses as possible for a common object, such as a brick or a paperclip. This test is particularly suitable for evaluating creativity because it captures both the fluency and originality of responses. In our study, we decided to focus on the AUT as it provides a clear and structured way to assess the creative outputs of LLMs. Understanding how these models perform on the AUT can offer insights into their potential and limitations in generating novel ideas.

Assessing creativity presents several challenges, primarily due to its subjective nature. Traditional methods involve human raters evaluating the originality and usefulness of responses, which can be time-consuming and prone to bias. To address these limitations, automated methods such as semantic distance measures have been developed, which quantify the relatedness of ideas to a given prompt. More recently, fine-tuning LLMs on human-judged responses has shown promise in providing reliable and scalable assessments of creativity. Each of these approaches has its strengths and weaknesses, highlighting the complexity of accurately measuring creative output.

In this research project, we aim to answer the following research question:

If and how can LLMs assist students to be more creative?

To this end, we first designed a comprehensive benchmark and an automated evaluation pipeline to assess the creativity of ideas generated on the AUT. This pipeline evaluates multiple dimensions of creativity, including originality, elaboration, dissimilarity, and flexibility. By establishing clear metrics and criteria, we aim to provide a robust framework for comparing human and LLM-generated responses. Secondly, we investigated various prompting strategies with different LLMs, ranging from zero-shot to few-shot prompting as well as different prompt structures to determine how the format and content of prompts influence the creativity of the generated ideas. By systematically varying the prompting conditions, we aim to identify the most effective strategies for fostering creative responses from LLMs.

1.2 RELATED WORK

In their seminal paper, Stevenson et al. 2022 explore the creative capabilities of GPT-3 by subjecting it to Guilford's Alternative Uses Test (AUT). This study critically examines GPT-3's ability to produce creative responses that are original, useful, and occasionally surprising. By comparing GPT-3's performance with human responses, the researchers found that, overall, humans outperform GPT-3 in creative output. However, the study suggests that with continued development, GPT-3 could potentially match human creativity. This research underscores the dual nature of LLMs as both impressive and limited, laying the groundwork for our investigation into leveraging LLMs for educational creativity

In this paper, Beaty and Johnson 2021 address the inherent limitations of subjective scoring in creativity research—namely, the labor cost and subjectivity—by introducing a method for automated scoring using semantic distance. The authors compare various semantic models and demonstrate that a latent semantic distance factor can reliably predict human creativity ratings. This work is crucial for our study as it validates the use of automated methods in assessing creativity, providing a robust framework for evaluating the outputs of LLMs on the AUT. The findings highlight the importance of objective measures in creativity research, which we incorporate in our methodology.

Organisciak et al. 2023 present a novel approach to automated scoring of divergent thinking tasks, surpassing traditional semantic distance methods. By fine-tuning LLMs on human-judged responses, they achieved a high correlation with human raters, significantly improving the reliability of creativity assessments. This research is directly relevant to our study, as it showcases the potential of advanced LLMs in accurately scoring creative outputs. The paper's insights into the limitations of purely semantic approaches and the benefits of fine-tuning inform our approach to comparing the creative performance of different LLMs and prompts.

Chapter 2

Methods

2.1 Dataset

2.1.1 Origin and preprocessing

We took the dataset from Organisciak et al. 2023 that originally is a collection of 9 datasets with approximately 27 000 samples from 3 different creativity tasks (uses, instances, consequences), with originality scored manually by humans (see figure 5.1 in Appendix A for more information on the original dataset). We managed to collect only 7 of those datasets and we kept only the samples applying to the Alternative Uses Test (type *uses*). After removing irrelevant columns, we ended up with a dataset of 14 584 samples, each one with an object associated (called *prompt*), a response generated by a human and its originality scored by a human too.

Note that during the data collection, humans doing the AUT were asked the following question for a given object: "What is a surprising use for a [NAME OF OBJECT]?".

2.1.2 Exploratory Data Analysis

After collecting all the data we could from the work of Organisciak et al. 2023, a quick exploratory data analysis (EDA) has been performed (see Fig 2.1).

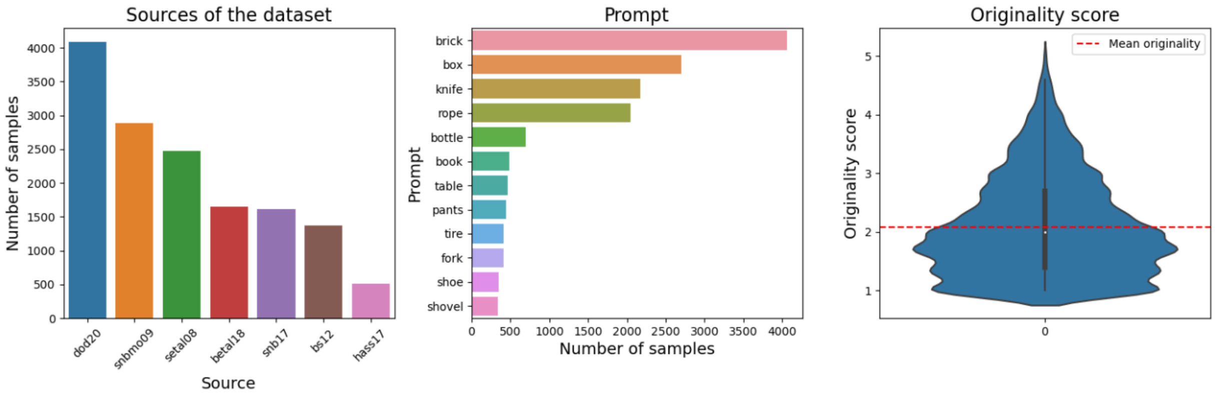


Figure 2.1
EDA dataset

From the EDA, we can see that the sources of the dataset are not really uniform but their originality distributions are pretty similar as seen in Fig. 2.2 so that we can reasonably combine all the datasets together without having to normalize each dataset. Looking at the distribution of objects in the dataset, it has been chosen to only keep the 4 most present objects, namely *brick*, *box*, *knife* and *rope*. Also, one can see that most of the samples have an originality around 2 (on a scale of 0 to 5), with few samples being very original (above 4).

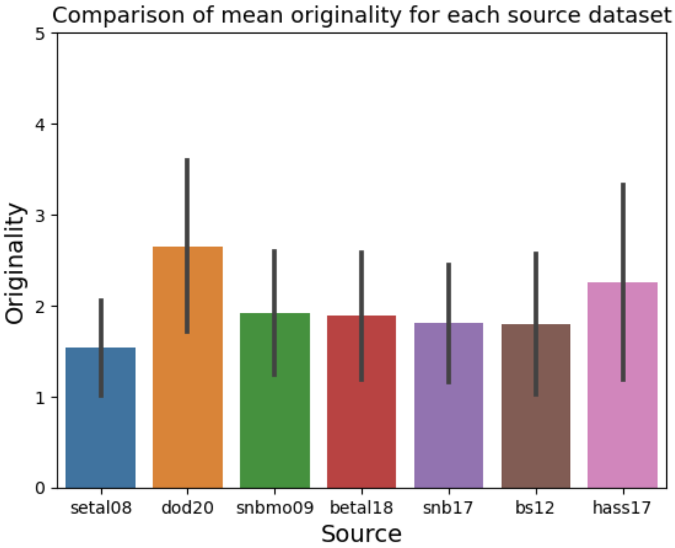


FIGURE 2.2
Comparison of originality of source datasets

2.2 MODELS

In this project, we employed a variety of large language models (LLMs) to explore and assess their capabilities. Primarily, we utilized OpenAI's GPT-3.5 (specifically *gpt-3.5-turbo-0125*) and GPT-4 (specifically *gpt-4-0125-preview*) models, accessing them through the OpenAI API. These models are renowned for their advanced natural language processing abilities, offering state-of-the-art performance across a wide range of tasks.

In addition to the proprietary GPT models, we experimented with several open-source large language models (LLMs) including Mistral 7B[1], Vicuna 13B[2], and Llama2 7B[3]. These models were run on a local server to mainly manage computational requirements. In fact, the decision to use a local server was driven by the significant hardware demands of these models, with most requiring at least 16GB of RAM for effective operation.

2.3 BENCHMARKING CREATIVE IDEAS ON THE AUT

2.3.1 OVERVIEW

Creativity is a multifaceted construct that can be evaluated across several dimensions. In the context of the Alternative Uses Test (AUT), a widely used measure of creative thinking developed by J.P. Guilford, four primary dimensions are commonly assessed: fluency, originality, flexibility, and elaboration. These dimensions offer a comprehensive evaluation of creative output, providing a robust framework for benchmarking creativity.

[1] `https://huggingface.co/mistralai/Mistral-7B-v0.1`
[2] `https://huggingface.co/TheBloke/Wizard-Vicuna-13B-Uncensored-GPTQ`
[3] `https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`

- **Fluency**: refers to the total number of alternative uses generated by an individual for a given object.
- **Originality**: measures the uniqueness or novelty of the responses provided.
- **Flexibility**: refers to the breadth of categories or conceptual spaces covered by the ideas generated.
- **Elaboration**: measures the amount of detail provided in the responses.

In our context of comparing the creative outputs of LLMs, we will not use the fluency dimension because it is a variable we can directly control when prompting the LLMs and would not be a relevant basis of comparison between LLMs as well as between LLMs and humans. However, we decided to add another dimension that we called *Dissimilarity* to measure the semantic diversity among the ideas generated by a LLM (more details in subsequent section).

Goal with scalability considerations in mind

One primary application of this project is in educational settings, where it can be used to assist a large number of students to be more creative during a brainstorming session. In the case where the task requires to be creative, it would be impractical and time-consuming for the teaching staff to manually score each response and provide a personalized feedback to each student. With this context in mind, we designed a creativity benchmark that is fully automated ensuring that creativity assessments are scalable, consistent, and objective. By automating the evaluation process, educators can efficiently manage large volumes of responses without compromising the quality of the assessment and gather valuable insights into the creative processes of students. By leveraging advanced NLP techniques and automated scoring systems, this project aims to create a robust and scalable solution for evaluating creativity on the AUT and hopefully for other tasks.

In the following sections, each dimension will be explained in detail and how they have been implemented in the evaluation pipeline.

2.3.2 ORIGINALITY

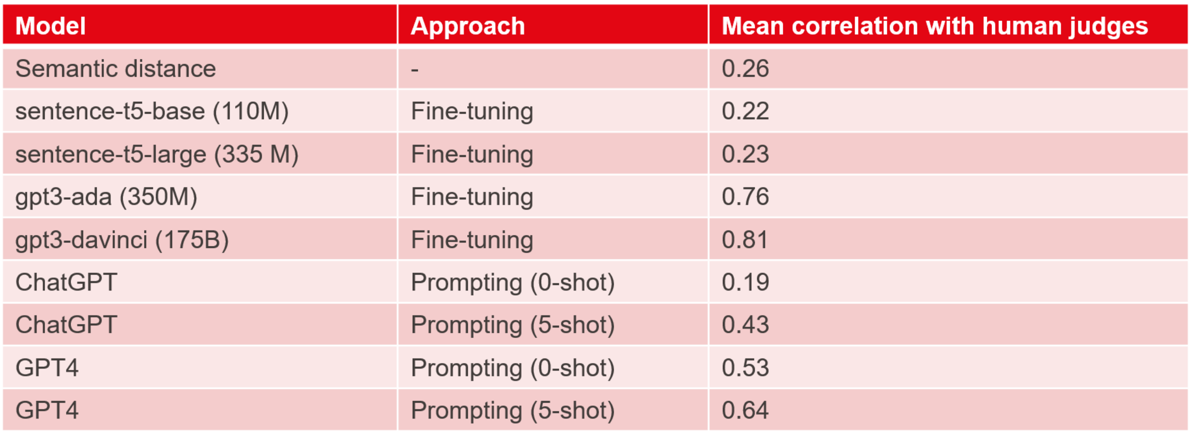
Originality measures the uniqueness or novelty of the responses provided. This dimension assesses how rare or uncommon the ideas are, highlighting the individual's capacity for innovative thinking. In creativity research, originality is often considered the hallmark of creative thinking (Runco and Jaeger 2012). Responses are typically scored based on their rarity among a given set of answers.

Originality Assessment

Scoring originality faces several challenges including reliability and scalability limitations. Recent developments in Natural Language Processing (NLP) techniques allow to automate the creativity scoring of ideas, removing the need to have a large number of human judges, all with their own subjectivity.

Several approaches exist to automatically score originality with different levels of complexity ranging from semantic distance scoring (Beaty and Johnson 2021) to AI-based scoring (Organisciak et al. 2023). Organisciak et al. 2023 fine-tuned deep neural network-based LLMs on human-judged responses and achieved up to $r = 0.81$ correlation with human raters, surpassing by far other approaches like semantic distance scoring ($r = 0.26$).

In this project, we will thus rely on OCSAI (Open Creativity Scoring with Artificial Intelligence), a tool designed by Organisciak et al. 2023, to automatically score the creativity of any idea. The evaluation simply consists of API calls with the name of the object and the idea as inputs. Specifically, we will use the *ocsai-chatgpt* model, a "GPT-3.5-size chat-based model, with slower scoring but slightly better performance than ocsai-davinci2", which achieved $r = 0.81$ correlation with human raters (see Fig. 2.3).



Model	Approach	Mean correlation with human judges
Semantic distance	-	0.26
sentence-t5-base (110M)	Fine-tuning	0.22
sentence-t5-large (335 M)	Fine-tuning	0.23
gpt3-ada (350M)	Fine-tuning	0.76
gpt3-davinci (175B)	Fine-tuning	0.81
ChatGPT	Prompting (0-shot)	0.19
ChatGPT	Prompting (5-shot)	0.43
GPT4	Prompting (0-shot)	0.53
GPT4	Prompting (5-shot)	0.64

FIGURE 2.3
OCSAI results

2.3.3 FLEXIBILITY

Flexibility refers to the breadth of categories or conceptual spaces covered by the ideas generated. This dimension evaluates the ability to shift between different categories or perspectives, demonstrating the individual's cognitive flexibility. Nijstad et al. 2010 emphasize that flexibility in idea generation is crucial for adaptive and innovative problem-solving, as it allows individuals to approach problems from multiple angles. In their paper, the authors wrote: "One prediction is that creativity benefits from an intellectual and dispositional capacity to generate and use remote associations". In the context of the AUT, we pose that the number of topics covered by an idea is a measure of the flexibility.

Flexibility Assessment

To measure this dimension, we employed a topic modeling algorithm on the dataset explained in the previous section, serving as our ground truth. This approach allowed us to extract distinct topics for each object and evaluate the flexibility of ideas generated by both humans and LLMs (see Fig. 2.4).

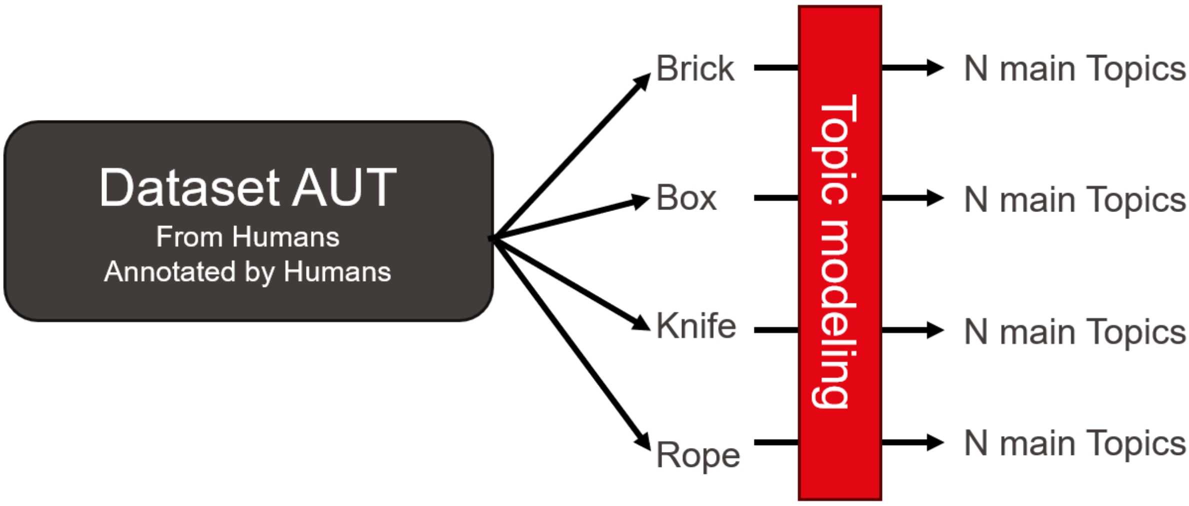


FIGURE 2.4
LDA pipeline

Our approach followed the following steps:

1. Choosing the Topic Modeling Algorithm

There exist a multitude of topic modeling algorithms with different approaches and different degree of complexity. For instance, Latent Semantic Analysis (LSA) is a topic modeling algorithm based on singular value decomposition of the TF-IDF matrix and assumes words with similar meanings will appear in similar documents. Another widely used topic modeling algorithm is Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003), a probabilistic approach that treats all words in a bag of words and assumes that each document is a combination of a small number of latent topics and each topic is a combination of keywords. Conceptually, LSA seeks to discover the underlying relationship between words while LDA seeks to discover the underlying topics in a corpus of text. LDA thus aligns better with our need as we would like to retrieve the main topics after combining all the creative ideas on one given object, namely retrieve the topics from a bag of creative ideas.

LDA is an unsupervised machine learning algorithm and similar to the clustering algorithm K-means, it will attempt to group words and documents into a predefined number of clusters (i.e. topics), that we need to initially set. Typically, once the number of topics is provided to the algorithm, all it does is to rearrange the topic distribution within documents and key word distribution within the topics to obtain good composition of topic-keyword distribution. The number of topics is thus a strategic parameter for a qualitative topic modeling algorithm.

2. Determining the Number of Topics

In LDA, each topic is a distribution over words. Typically, the N most probable words per topic represent that topic. The idea is that if the topic modeling algorithm works well, these top-N words are semantically related. The difficulty is how to evaluate these sets of words. To this end, we will use the coherence, a widely used metric to evaluate a topic modeling algorithm. Briefly, it measures the degree of semantic similarity between high-scoring words in a topic and helps evaluate whether the topics generated are interpretable and meaningful for humans. You can find more information about the computation of the coherence in the Appendix 5.2.

To determine the optimal number of topics, we thus analyzed the coherence plot for each object by varying the number of topics.

To this end, we first gathered all the ideas of the same object together and applied some preprocessing functions (remove stop words, punctuation and apply lemmatization). Then, for each object, we trained a LDA model by passing all the ideas of the corresponding object and evaluated its coherence score.

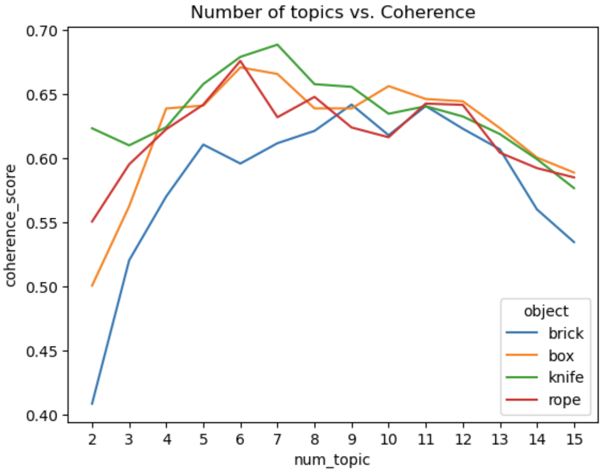


FIGURE 2.5
Coherence score vs number of topics for each object

After analyzing the coherence plots, we found that using 6 topics provided the most coherent and distinct topics for each object in average.

3. Determining the Optimal Number of Keywords

Our context of benchmarking creativity along several dimensions requires to give a flexibility score to each idea that will be generated. To this end, each evaluated idea will be assigned a score that represents the number of topics covered by the idea. LDA provides a list of keywords, with varying importance weights, for each topic. The computation of the flexibility score thus involves assigning a score to each sentence based on the presence of specific keywords in the sentence. Here's the detailed process:

1. **Topic Modeling:** For each object, LDA is used to generate 6 topics, each characterized by N keywords.
2. **Sentence Evaluation:** Each idea is evaluated against the list of keywords for all topics.
3. **Keyword Matching:** If a sentence contained a keyword from a topic, it is assigned to that topic.
4. **Flexibility Scoring:** Each sentence assigned to a topic increased the flexibility score by 1 point.

To compute the flexibility score of any given idea, it was essential to determine the optimal number of keywords for each topic. In fact, the number of keywords impacts the robustness and accuracy of the topic assignments. If too few keywords are used, many ideas may receive a flexibility score of zero because no word will be matched. Conversely, using too many keywords can lead to overlap, where one word in a sentence being the keyword of several topics will not account for a relevant flexibility score. The goal is thus to maximize the number of keywords to ensure broad topic coverage while minimizing keywords overlaps between different topics.

To strike a balance, we examined the keywords overlaps from the LDA with varying numbers of keywords. Checking the overlap of keywords generated by the LDA might not be intuitive as LDA is trained, by design, to rearrange the distribution of keywords to have a distinct set of keywords for each topic. However, we noticed that for a given number of topics and a given number of keywords per topic, this rule was not respected. For example, we started to see some overlaps with 6 topics and 30 keywords (see in Appendix Fig 5.2). Interestingly, for 7 topics, the overlap started at 10 keywords per topic (see in Appendix Fig 5.3).

After analysing the overlaps of keywords for 6 topics, 20 keywords per topic appeared as a good trade-off, with a reasonable number of keywords to ensure broad topic coverage and no overlap as seen in Fig 2.6.

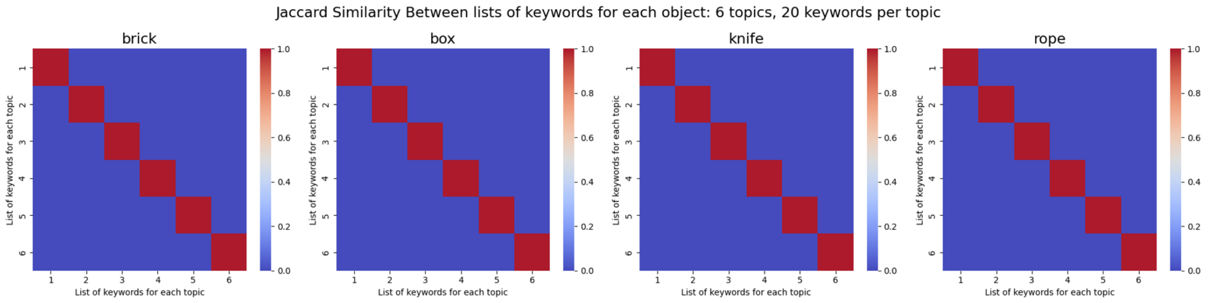


FIGURE 2.6
Overlap of keywords with 6 topics and 20 keywords per topic

FLEXIBILITY AUGMENTED Building upon this flexibility score, it has been decided to design an augmented flexibility score that could account for the breadth of topics covered and the originality of those topics, namely combine the flexibility and the originality dimensions.

This augmented flexibility score for one sentence is computed as follows:

$$\text{Score flexibility augmented (sentence)} = \sum_{t=1}^{T} N_t \cdot k_t$$

with:
N_t : Number of words in the sentence matching the keywords of topic t of a given object..
k_t: Coefficient associated with topic t of a given object.

The idea behind this score is that several topics might not have the same originality. We suppose that this augmented score will allow to shed lights on topics that make an idea original. Said differently, this score computes a flexibility score but adds more weight on the topics that rarely appear in the humans dataset. The weights come from the coefficients associated to each topic for each object. The computation follows these steps:

1. **Topic Modeling:** For each object of the humans dataset, LDA is used to generate 6 topics, each characterized by N keywords.

2. **Topic assignment:** Each idea of the humans dataset is assigned a unique topic by comparing its words to the keywords of each topic for the corresponding object.

3. **Coefficient computation:** Each topic receives a coefficient equals to 1 minus its frequency in the topic assignments. Said differently, for each topic of each object, we first divide the number of sentences assigned to topic t by the total number of sentences in the dataset, that we call the frequency of appearance in the dataset, and we substract this score to 1, namely 1 - frequency. Consequently, a topic rarely appearing, namely that has been assigned very few times to ideas, will have a very low frequency and thus a very high coefficient so that a sentence covering this topic will have a high flexibility augmented score.

This method required to check the number of keywords again. In fact, we again performed a unique topic assignment. This means that for each sentence, we iterate over all topics, and iteratively we check whether the sentence has a word from the keywords of the given topic. If there is match, the sentence is assigned to this topic, otherwise, we iterate through the next keywords' topic. Remember that we check the overlap of keywords generated by the LDA model but this does not guarantee that one sentence does not have a word from the keywords of topic A and one word from the keywords of topic B. In this scenario and because of how we iterate through the topics' keywords for the topic assignment, a sentence can be assigned to several topics and the final topic assignment will default to the last match found.

This can clearly be seen in Fig. 2.7 on the right side with 50 keywords. One can see that last topics (4 and 5) are consistently assigned to more sentences than the first topics. This might be due to the fact that one sentence has words from several topics. One can see that the same problem occurs for 20 keywords per topic (see left side of Fig 2.7).

To counter that, we are looking for an optimal number of keywords that does not show this pattern of increasing assignments of topics across all objects like in Fig 2.7. Looking at Fig 2.8, we see that 10 keywords per topic satisfy this condition and we decided to take N = 10 keywords per topic for all flexibility scores that will be done in this project, including the original flexibility score previously explained for consistency reasons.

All in all, the flexibility assessment consists in counting the number of topics covered by an idea (by considering 6 topics and 10 keywords by topic) and the augmented flexibility assessment consists in counting the number of topics covered and weighting them by the coefficient associated to the topic.

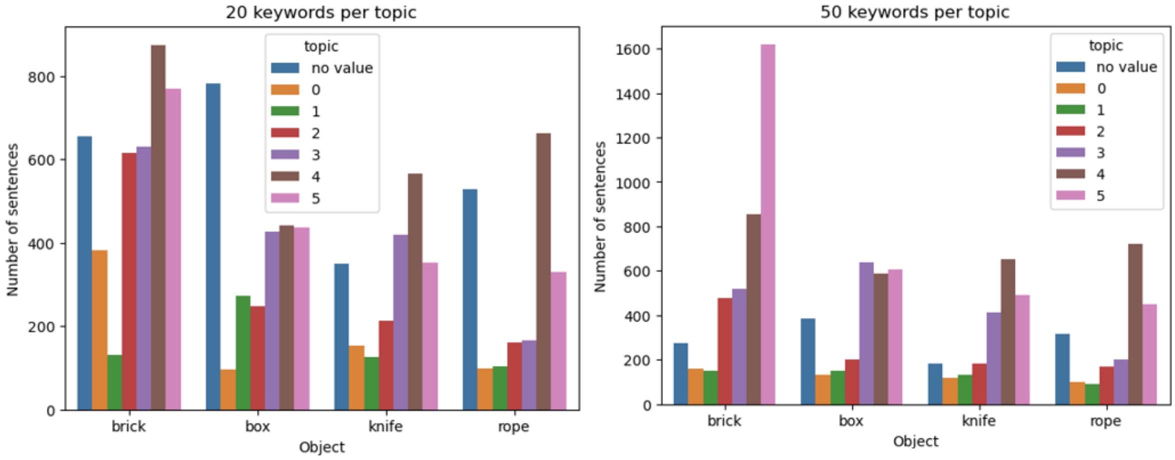


FIGURE 2.7
Assignment of sentences with varying number of keywords

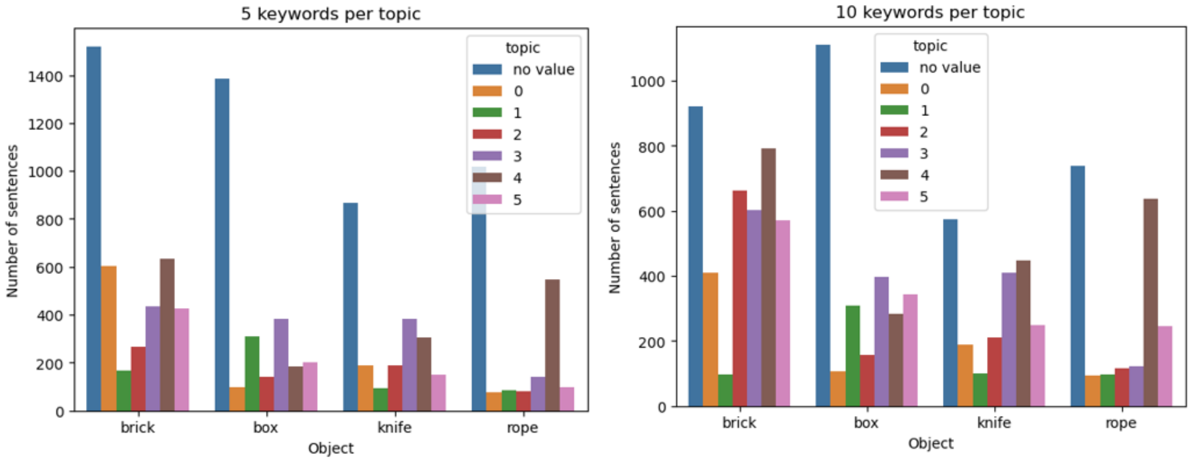


FIGURE 2.8
Assignment of sentences with varying number of keywords

2.3.4 DISSIMILARITY

The dissimilarity, computed as $1 -$ similarity, aims to measure the disparity of each idea proposed for a given object. In fact, we prefer a large language model (LLM) that generates a diverse set of ideas with varying originality over one that produces highly original but almost similar ideas.

Dissimilarity assessment

The dissimilarity assessment is inspired from the BERTScore (Zhang et al. 2019) except that we compute the cosine similarity on the sentence embeddings and not the words embeddings. Note that this twist came from the realization that BERTScore did not seem to align perfectly with our use case. In fact, BERTScore is typically useful for evaluating the quality of text summarization, measuring how similar the text summary is to the original text. In our case, the texts being compared are relatively short sentences and we suppose that averaging all the words' embeddings to have the sentence embedding keeps its relevance.

To this end, we leverage the capacities of a pretrained embedding model like BERT, specifically the *distilbert_base_uncased* [4] model for computational reasons, to first compute the contextual embeddings of each word of the idea being evaluated. We then pooled (mean pooling) all the words' embeddings together to create a single representation of the sentence (in this case each sentence is represented by an embedding vector of 768 elements). For a given sentence, we first compute the cosine distance of its embedding vector with all the other sentences' embeddings, for the same object. We then average this score and substract it to 1 to obtain the dissimilarity (see Fig 2.9).

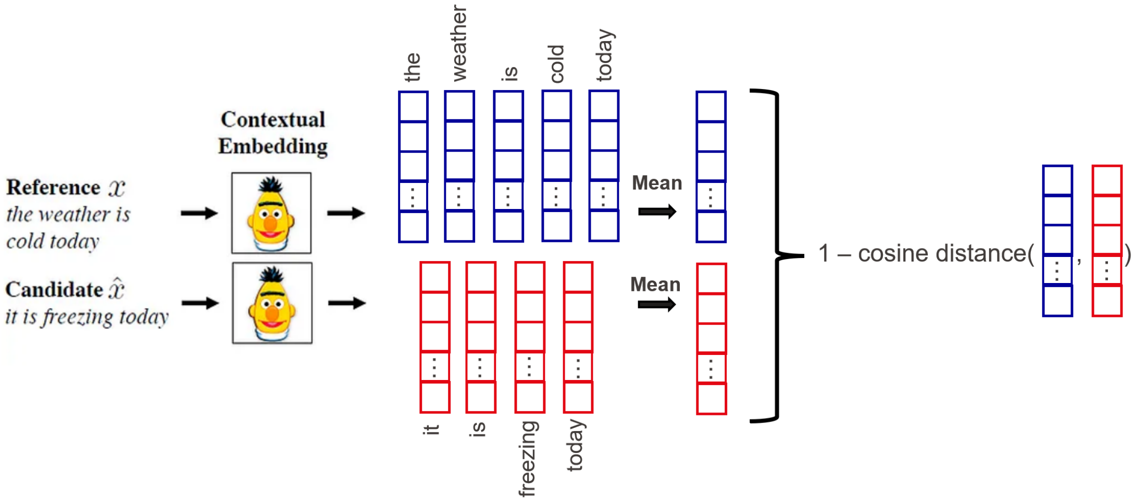


FIGURE 2.9
Dissimilarity assessment

2.3.5 ELABORATION

Elaboration measures the amount of detail provided in the responses. This dimension reflects the depth and richness of the ideas, indicating the individual's ability to develop and expand upon initial thoughts. Elaboration is linked to the capacity for detailed thinking and the ability to extend simple ideas into more complex and well-defined concepts.

Elaboration Assessment

Three approaches have been explored to evaluate the richness of the responses, with an increasing degree of depth. First, the simple elaboration consisted in counting the number of words for each idea, reflecting directly the amount of detail provided. Secondly, an elaboration without the stop words has been implemented to remove words that did not bring a lot of detail in the idea. Ultimately, we wanted to design an elaboration score based on Part-of-Speech tagging. In fact, a rich sentence often includes a diverse set of parts of speech, such as nouns, verbs, adjectives, and adverbs. For example, sentences with many descriptive adjectives and adverbs are usually more detailed. Instead of directly counting the number of words, this elaboration score would consist in counting the number of nouns, verbs, adverbs, adjectives for instance.

Unfortunately, the last two approaches did not achieve promising results and we kept the simple elaboration score as the number of words in the idea (see more detail in Appendix 5.4). The elaboration was actually computed directly by OCSAI, at the same time than the originality, by setting the parameter *elab_method* to *whitespace* when calling the OCSAI API.

[4] https://huggingface.co/distilbert/distilbert-base-uncased

2.4 LLMS PROMPTING STRATEGIES

Prompt engineering is a crucial process in the utilization of large language models (LLMs). It involves crafting specific inputs, or prompts, to guide these models in generating desired outputs. The effectiveness of a LLM is significantly influenced by the quality of the prompts provided. Prompt engineering aims to maximize the model's performance by optimizing these inputs.

2.4.1 OVERVIEW OF PROMPTING STRATEGIES

There are several prompting techniques used in practice such as (among others):

1. **Zero-shot prompting:** This technique involves providing the model with a task description without any examples. The model relies on its pre-trained knowledge to generate an output. Zero-shot prompting is flexible but depends heavily on the model's pre-existing knowledge base.
2. **One-shot prompting:** This method provides the model with a single example along with the task description. It helps the model understand the specific nature of the desired output by showcasing an example of the correct response.
3. **Few-shot prompting:** By presenting a few examples, this approach allows the model to better grasp the pattern and nuances of the task, leading to more accurate and contextually appropriate outputs.
4. **Chain-of-thought prompting:** This technique involves guiding the model through a series of logical steps to reach the final answer. Breaking down complex tasks into smaller, manageable steps helps the model produce more coherent and logical outputs.
5. **Contextual prompting:** In this strategy, additional context or background information is provided to help the model generate more relevant and informed responses. This can include detailed descriptions, historical data, or situational specifics that shape the model's understanding and output.

2.4.2 INFLUENCING FACTORS

Prompt engineering is influenced by several key factors that determine the quality and relevance of the generated outputs. In this project, we decided to focus on the two following factors: the general framework of the prompt and several few-shot prompting approaches.

Framework

The structure of a prompt is critical in guiding the model's output. Effective prompts can typically be structured to include the following elements: persona, context, task, format, and examples. This framework ensures that the model has all the necessary information to generate high-quality and relevant responses.

1. **Persona:** Defining a persona involves specifying the identity or character the model should adopt while generating the response. This influences the tone, formality, and style of the output. For example, instructing the model to respond as an experienced software engineer or a casual friend can tailor the response to match the desired communication style.
2. **Context:** Providing the necessary background information or situational details helps the model understand the environment or scenario. Contextual information can significantly enhance the relevance and accuracy of the model's responses. For instance, including information about recent events or specific settings can guide the model to generate more contextually appropriate outputs.

3. **Task:** The task defines the specific objective or function the prompt is intended to achieve. A clear articulation of the task ensures that the model's output aligns with the desired goal. Each prompt should begin with a precise action verb and goal.
4. **Format:** The format dictates the structural arrangement or layout of the output. Specifying the format ensures that the response meets structural requirements and is organized appropriately.
5. **Examples:** Examples or samples included in the prompt serve as guides for the model's output. These examples establish a benchmark for the desired response style or structure, ensuring consistency and adherence to specific formats. Effective exemplars can significantly improve the quality of the generated outputs.

In our case, we first designed an initial prompt for the GPT models (specifically *gpt-3.5-turbo-0125* and *gpt-4-0125-preview*) as shown below under **Initial prompt**. This prompt was carefully crafted to achieve several key objectives. First, it clearly defines the task by explicitly instructing the model to "Generate exactly N_responses alternative uses for the object [object_name]." This ensures the model's output is aligned with the specific goal. Second, it imposes a reality constraint by emphasizing ideas that are "useful in real life" and "while staying realistic," which encourages practicality. Third, to prevent the usual lengthy and verbose responses typical of LLMs, the prompt mandates that each idea be "a concise sentence." Fourth, the format is standardized by requiring each alternative use to follow the given examples, thus maintaining consistency. Note that for this part of the experiment, we always used the same examples (*Sock, Color it and maybe make a snake* and *Sock, Use it as a puppet*) from an object that's not one of the four objects tested to avoid any influence. Lastly, the prompt encourages originality by highlighting the need for "creative, out-of-the-box ideas" that are "especially appreciated if they are original." This structured approach ensures that the generated ideas are not only diverse and innovative but also feasible and relevant to real-world applications.

Then we tested the effectiveness of our prompt engineering strategy by evaluating various modifications of this initial prompt. This allowed us to understand the impact of each constraint on the quality and creativity of the generated responses. Our testing strategy included the following variations:

Initial prompt without creative constraint (P2): This version removed the emphasis on proposing "realistic" ideas that should also be "useful in real life", allowing us to see how the model performed without being explicitly pushed towards realistic ideas.

Initial prompt without length constraint (P3): Here, we eliminated the requirement for each idea to be "a concise sentence." This helped us assess whether the length constraint was effective in generating focused and succinct responses and whether it impacted the creativity of the outputs.

Initial prompt without persona (P4): By removing the persona aspect ("You are a very creative, open-minded person"), we aimed to determine how important the persona is in guiding the model's creativity and open-mindedness.

Initial prompt without persona and context (P5): This variation stripped away both the persona and the context ("You are meant to assist students in group ideation..."), allowing us to evaluate the combined effect of these elements on the model's output.

Prompt like humans (P6): This variation wanted to get as close as possible to the question asked to humans in the collection of the dataset used by Beaty and Johnson 2021 and ourselves. Specifically, the humans were asked "What is a surprising use for [name of the object]?", which has been included in the prompt. Though the task and the format parts have been kept to maintain consistency in the outputs compared to other prompts.

Initial prompt without creative and length constraints (P7): In this version, both the creative constraint and the length requirement were removed. This allowed us to observe how the absence of these two critical components affected the diversity and practicality of the generated ideas.

By systematically testing these variations, we aimed to identify which components of the prompt were most critical in producing high-quality, creative, and useful alternative uses for objects. This iterative approach provided insights into the importance of each prompt element and informed our prompt engineering practices. Also we aimed to investigate whether these variations have a different effect depending on the LLM (GPT-3.5, GPT-4, open-source models,...). Some of the prompts can be seen in their detailed form right below (P1, P2 and P6) and the other prompts can be seen in Appendix 5.4.

FRAMEWORK

Initial prompt (P1)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Initial prompt without creative constraint (P2)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Prompt like humans dataset (P6)

What is a surprising use for a object_name?
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Few-shot examples

As mentionned by Brown et al. 2020, LLMs are few-shot learners and providing examples in the prompt can significantly enhance their performance. In our case, besides first using a few examples to format the output, we were wondering to what extent more fine-grained few-shot strategies could help the LLM generate creative ideas.

In a typical educational context, few-shot prompting could be used for human-LLM collaboration. Namely, students would be first asked to generate creative ideas and then we could provide their ideas to the LLM via the few-shot examples and invite the LLM to explore other topics.

For this part of prompt engineering, we took as baseline the initial prompt (P1) and analyzed 5 different configurations as listed below:

1. **5 fs Max:** 5 few-shot examples of the corresponding object carefully selected in the humans dataset were provided in the prompt. For each object, we first ranked all the ideas of the humans and selected the top 5 based on the originality dimension.
2. **5 fs Random:** 5 few-shot examples of the corresponding object **randomly** selected in the humans dataset were provided in the prompt.
3. **5 fs RS:** 5 few-shot examples of the corresponding object **randomly** selected in the humans dataset with their corresponding originality **score** were provided in the prompt.
4. **10 fs RS:** 10 few-shot examples of the corresponding object **randomly** selected in the humans dataset with their corresponding originality **score** were provided in the prompt.
5. **20 fs RS:** 20 few-shot examples of the corresponding object**randomly** selected in the humans dataset with their corresponding originality **score** were provided in the prompt.

Note that for each configuration, we added a condition that each example must have at least 5 words. In fact, we noticed that some ideas in the humans dataset had 1-2 words and the maximum originality of 5, which might be an outlier that could wrongly influence the LLM to output one word ideas.

Again we aimed to identify which few-shot variations were most critical in producing creative alternatives uses for objects and whether these variations have a different effect depending on the LLM (GPT-3.5, GPT-4, open-source models,...). The few-shot prompts used can be seen in their general form right below (called **N fs Max** and **N fs RS**) and specific examples with real values can be seen in Appendix 5.4.

PROMPTS FEW-SHOT

Prompt few-shot max (N fs Max)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below: fs_examples

Prompt few-shot random with scores (N fs RS)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to proposealternative uses for an object, and you should share your ideas of alternative uses to inspire them to exploreother possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below: fs_examples
Below are the originality scores for each example listed in order. The scores range from 0 to 5. Use these scores to understand what makes an idea original, but do not include them in your output.
Originality scores: fs_scores

Chapter 3

Results and Discussion

3.1 Comparing the creative outputs of GPT models

In all the results presented, GPT models were prompted to generate 100 ideas for each object, resulting in approximately 400 ideas per object. In practice, we rarely achieved exactly 400 ideas per object, typically obtaining between 350 and 400 ideas, as the LLMs often did not strictly respect the specified number of alternative uses. Using 100 alternative uses strikes a balance between generation time, cost, and representativeness. Systematically, the different prompts will be compared to the alternative uses generated by humans.

3.1.1 Univariate analysis

Comparison 1: framework

This first part of the analysis focuses on the different framework configurations of prompts (P1 to P7) and are based on the results shown in Fig 3.1.

In terms of distribution density: It is interesting to note that for both GPT-3.5 and GPT-4, the distribution density of originality and elaboration differs significantly from that of humans, with outliers present only in the human data. For the other three dimensions (flexibility, augmented flexibility, and dissimilarity), the distributions of outputs from both GPT models and humans are approximately the same. Specifically, for the flexibility scores of both humans and GPT models, the scores are concentrated around integers. However, this pattern is observed only for GPT models in the originality scores, where many ideas are condensed around an originality score of 2 or 3, while human ideas have a more uniform distribution from 0 to 3.

In terms of values, especially when comparing the medians: For both GPT-3.5 and GPT-4, the median scores for originality and elaboration are always slightly higher than those of humans. However, human ideas tend to have a higher median score for dissimilarity. This suggests that human ideas are rated as less original and detailed but have greater semantic diversity (dissimilarity). The median scores for the two flexibility metrics are the same across humans and both GPT models.

Of the seven prompts used, one stands out for both GPT-3.5 and GPT-4: prompt P3, which lacks a length constraint. This prompt is particularly notable for its originality, showing some outlier values, although the medians are similar to those of other prompts. For dissimilarity, prompt P3 has a slightly higher median. Interestingly, although this prompt was intended to generate longer outputs, the elaboration scores did not reflect this (the medians are all pretty similar for all prompts).

When comparing the outputs of GPT-3.5 and GPT-4, the results are generally similar. However, some outliers are apparent in the dimensions of originality and flexibility. Specifically, GPT-4 has outliers with higher values on the originality dimension for each prompt compared to GPT-3.5 in overall, and GPT-3.5 has higher outliers than GPT-4 on the flexibility scores overall.

Interpretation of Results:

Firstly, it is important to note that the human ideas analyzed come from seven different sources, potentially involving tens of individuals, whereas the ideas generated by the LLMs come from a single "entity." This diversity of human sources likely contributes to the higher dissimilarity values. Secondly, as deterministic models, LLMs tend to produce deterministic outputs, which counteracts the generation of a diverse set of ideas, especially when prompted to generate 100 ideas at once. For elaboration, although the LLMs were often prompted to provide each alternative use in a "concise sentence," their elaboration scores were still higher than those of humans. We hypothesize that this might be due to the inherent difficulty humans face in generating creative ideas with a lot of details, a task that is easier for an LLM, especially given its tendency to produce verbose responses. Regarding the originality dimension, the U-shaped distribution of the GPT models' outputs stands out compared to the more uniform distribution of human ideas. It seems that LLMs either generate ideas that are not very original (originality around 2) or moderately original (originality around 3), but they struggle to produce highly original, out-of-the-box ideas (at least according to OCSAI).

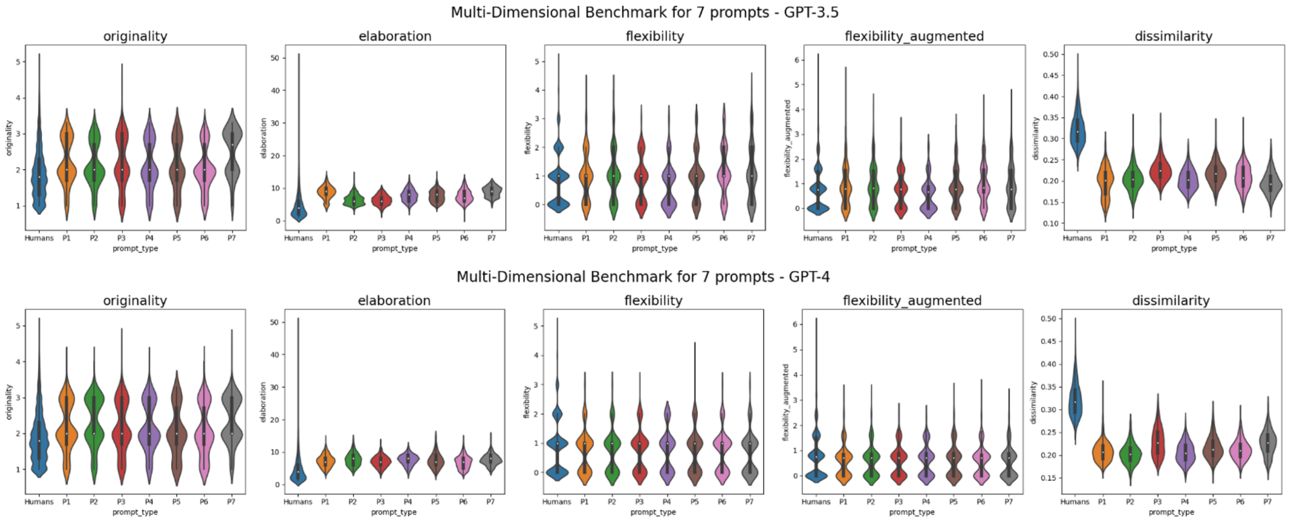


FIGURE 3.1
Prompts basic

Comparison 2: few-shot strategies

This second part of the analysis focuses on the different few-shot strategies tested and are based on the results shown in Fig 3.2. As a reminder, the different few-shot strategies are a variation of the initial prompt (P1) that we denote as *Baseline* in the the plots below.

Most of the conclusions drawn in the previous part are still valid. Specifically, Humans and GPT models' ideas have different distributions' densities for originality and dissimilarity; the originality and elaboration scores are overall higher for all few-shot approaches compared to Humans but their dissimilarity scores are lower (when looking at the medians); and the flexibility scores distributions are pretty similar in density and values between the GPT models and Humans, thought some prompts generated more outliers than others, especially when looking at GPT-3.5.

Of the 5 few-shot strategies tested, one stands out for both GPT-3.5 and GPT-4: prompt 5 fs Max, which consists in 5 few-shot examples carefully selected in the Humans dataset for having the highest originality. The ideas generated by this prompt particularly stand out for their originality, showing a high median (at 3) compared to other prompts and some higher outlier values. This prompt also stands out on the elaboration dimension for both GPT models. Further analysis could investigate whether the length of the provided examples impacted the LLMs outputs.

When comparing the variation of few-shot examples from maximum originality (5 fs Max) with the variation of randomly selected few-shot examples (5 fs Random), we can see that the originality seems to be always better overall for the first option (5 fs Max) when looking at the median, and for both GPT models. However the results seem to diverge when mentioning the scores of the randomly selected examples. In fact, when looking at the three variations of few-shot examples randomly selected with scores, that we denoted as "5 fs RS", "10 fs RS" and "20 fs RS", we can see that the impact of adding the information of the scores increased the median originality for GPT-4 for 5 and 10 few-shot examples. Nevertheless, this did not apply to GPT-3.5, though some outlier original ideas seem to have been generated (see the increased upper tail between 5 fs RS and 20 fs RS for GPT-3.5). Though the original plan behind the idea of adding scores to the few-shot examples (from 5 to 20) was for the model to learn what makes an idea original, it has been noted that in all three cases (for 5, 10 and 20 examples), the examples had an originality between 1 and 3 most of the time, preventing the model to see a diversity of original ideas and ideas covering the whole originality spectrum. Further experiment could consist in carefully selecting a diverse set of original ideas from not original to very original.

Interpretation of Results:

The few-shot strategies provide a deeper insight into the creative capabilities of GPT models compared to human ideation. The variation in the distributions of originality and dissimilarity between humans and GPT models is consistent with previous findings. GPT models tend to generate ideas with higher originality and elaboration scores but lower dissimilarity scores, indicating a tendency to produce more detailed and innovative ideas that are, however, less varied in nature.

The standout performance of the prompt "5 fs Max" strategy highlights the influence of high-quality few-shot examples on the models' outputs. By providing examples with high originality, the models are able to generate ideas that not only match but sometimes exceed the creativity levels seen in human ideas, as evidenced by the higher median originality scores and the presence of significant outliers. This suggests that few-shot learning can effectively guide the models towards producing more creative outputs.

However, the results obtained with the combination of ideas and originality scores were not very significant, though GPT-4 showed an increased originality and elaboration when adding the scores to the randomly selected few-shot examples. This experiment would require more advanced research to investigate whether adding more examples or adding more information (like the scores) has an impact on the outputs' creativity.

Overall, these results emphasize the importance of prompt design and examples selection in guiding LLM outputs. While GPT models can generate highly creative ideas, their performance is can be influenced by the quality (and apparently not the number) of the provided examples. This highlights the potential for tailored few-shot strategies to enhance the creative outputs of LLMs.

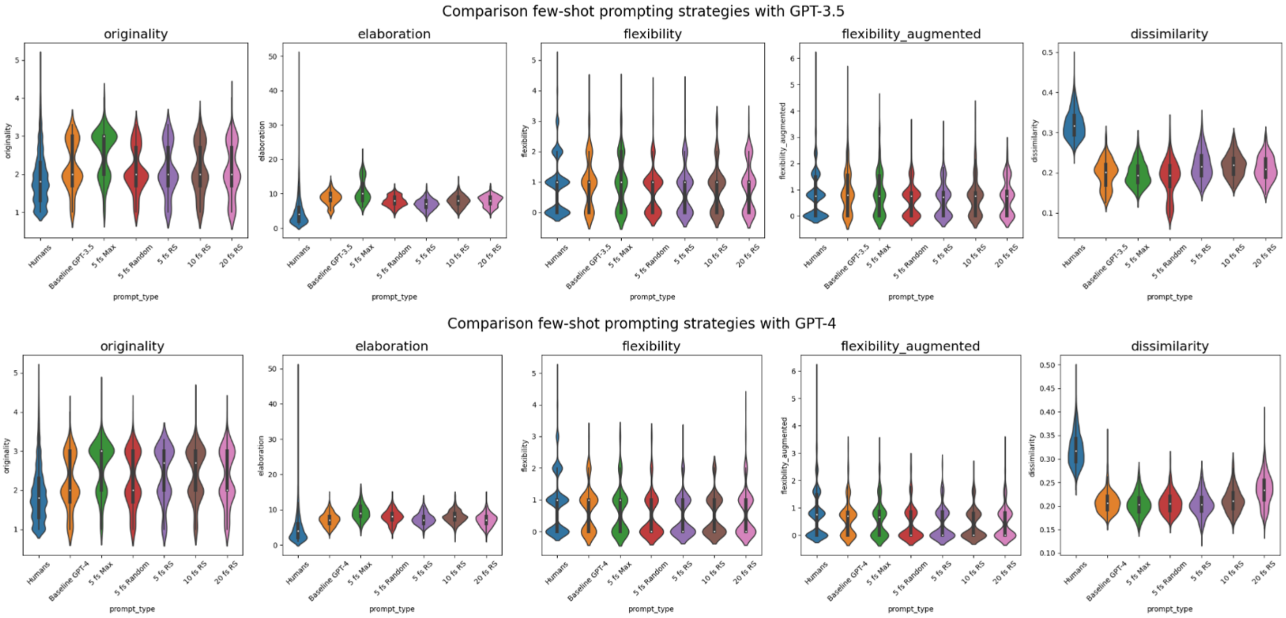


FIGURE 3.2
Prompts with few-shot examples

3.1.2 COMPARATIVE ANALYSIS BETWEEN OBJECTS

In this section, we compare the creativity dimensions per object for the framework prompts, focusing exclusively on GPT-3.5 and the originality dimension for the sake of conciseness.

When examining the originality dimension, as depicted in Fig. 3.3, it is noteworthy that humans are quite consistent, showing similar distributions and median values across different objects.

In contrast, GPT-3.5 does not exhibit the same consistency across all prompts, especially for P1, P2, P3, P4, and P5, as evidenced by the varying distribution densities and median values. However, GPT-3.5 demonstrates a more consistent performance across different objects for one of the two prompts closely aligned with the human task (P6): "What is a surprising use for ...?"

For the LLM-generated ideas, most objects display a U-shaped distribution in originality, with scores clustered around 2 and 3, except for the object "rope," which shows a more uniform distribution ranging from 1 to 4.

Additionally, the distributions across objects for prompt P3, which stood out in the previous analysis, reveal that it is primarily the object "brick" that significantly impacts originality.

Future research could investigate patterns across prompts and objects that positively influence creativity. Extending the between-objects comparison to GPT-4, other prompts variations and other creativity dimensions (elaboration, dissimilarity, flexibility) could provide further insights, with some analyses available in the Appendix (Fig. 5.5).

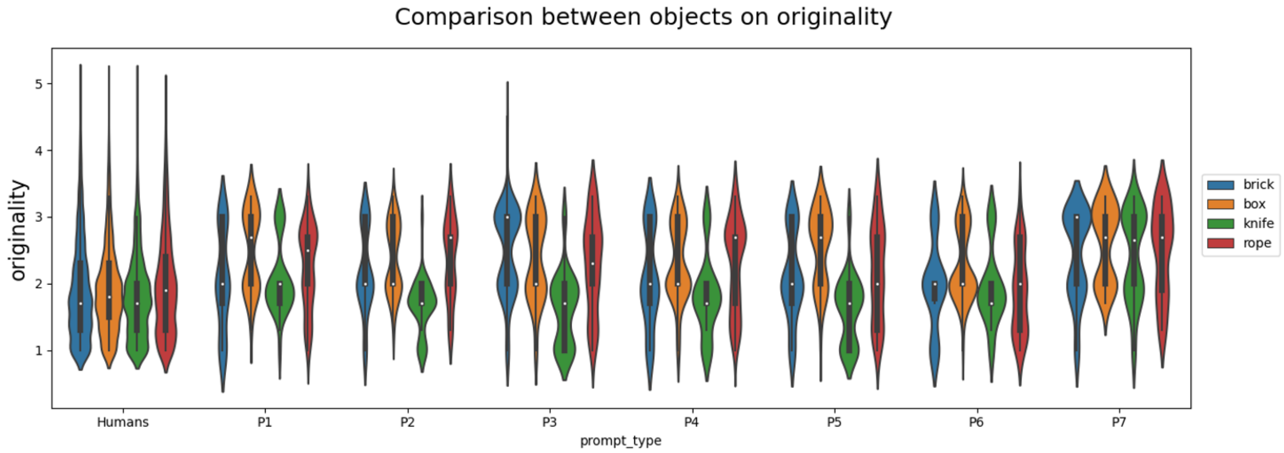


FIGURE 3.3
Comparison per object on GPT-3.5 outputs' originality for the "Framework prompts"

3.1.3 MULTIVARIATE ANALYSIS

In this section, we focus on the multivariate analysis of the creativity dimensions by examining the correlation heatmaps for different prompts. We selected the estimated best-performing prompts, P3 and 5 fs Max, for this comparison, while keeping humans and the initial prompt P1 as baselines and decided to focus only of ideas generated with GPT-4.

The correlation heatmaps (see Fig 3.4) provide insights into how different creativity dimensions interact with each other across various prompting approaches. By analyzing these heatmaps, we can draw conclusions about the relationships between features for each prompt and compare them to human responses. Ultimately, it could give us some insights into how humans generate creative ideas that could inform our prompting strategies with LLMs.

Similarities and Differences Between Humans and LLMs (for the prompts P1, P3, 5 fs Max)

The analysis reveals both similarities and differences in creativity dimensions between humans and the three prompts (P1, P3, 5 fs Max). All prompts, like humans, maintain a strong correlation between flexibility and flexibility augmented, which is expected since flexibility augmented is a derived measure. Elaboration consistently correlates positively with flexibility for humans and across all prompts, though it correlates a bit more for humans (0.55-0.57 for humans compared to 0.15-0.25 for LLMs).

However, some differences arise in the correlation patterns between humans and LLMs. First, originality correlates a bit more negatively with flexibility dimensions (normal and augmented) for LLMs (from -0.36 to -0.16) compared to humans (-0.05). Second, elaboration is positively correlated to dissimilarity for humans (0.41) but negatively correlated for LLMs (from -0.18 to -0.09). Practically speaking, it could mean that humans' verbosity was used to elaborate more diverse ideas while LLMs verbosity did not bring more diversity. Third, dissimilarity correlates negatively with flexibility scores for the LLMs (especially P1 and 5 fs Max) but the correlation is almost null for humans.

Similarities and Differences Between the prompts P1, P3, and 5 fs Max

When comparing the three prompts (P1, P3, 5 fs Max) among themselves, several patterns emerge. P1 and 5 fs Max share almost the same correlations patterns between the different dimensions. The only exception is the stronger positive correlation between originality and dissimilarity for 5 fs Max. However, humans and P3 share some common correlation patterns. Notably, dissimilarity does not correlate with the flexibility scores and they have a similar correlation between originality and dissimilarity (0.18 and 0.23).

Nevertheless, humans ideas exhibits a stronger correlation between elaboration and all other dimensions.

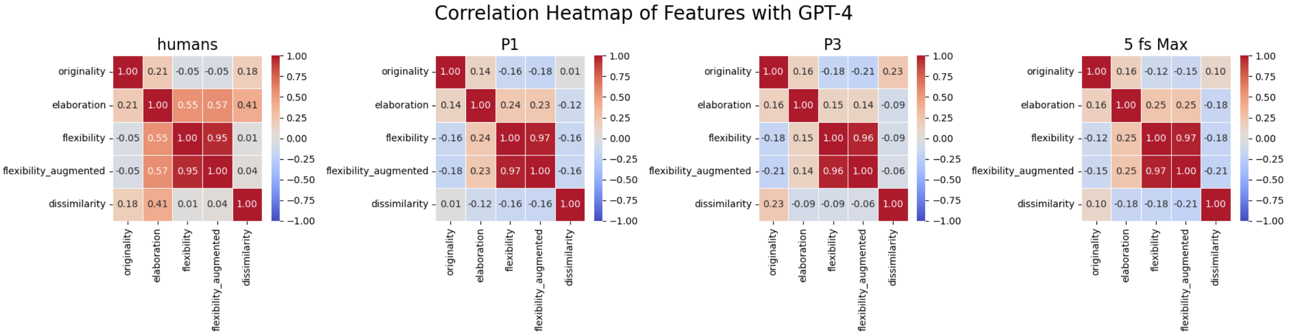


FIGURE 3.4
Multivariate analysis for Humans vs P1 vs P3 vs 5 fs Max with GPT-4

In summary, while GPT-4 model exhibit some similar correlation patterns to humans in terms of flexibility and elaboration, significant differences in originality and dissimilarity correlations highlight the distinct nature of machine-generated creative outputs. The nuanced differences between prompts suggest that different prompting strategies can significantly influence the relationships between creativity dimensions. Future research could explore these patterns further and investigate how specific prompting strategies can be optimized to enhance the diversity and originality of AI-generated creative ideas. Additionally, expanding the analysis to include GPT-3.5 could provide deeper insights into the capabilities and limitations of LLMs in generating human-like creative ideas.

3.2 THE CASE OF OPEN-SOURCE MODELS

In our study, we explored the use of several open-source large language models (LLMs) including Mistral 7B, Vicuna 13B, and Llama2 7B that we used on a local server. These models were selected for their potential to offer more cost-effective and privacy-respecting alternatives to proprietary models like GPT-3.5 and GPT-4.

In fact, utilizing open-source models can significantly reduce the financial burden associated with using the API of large proprietary LLMs like GPT models, especially in large-scale deployments. Additionally, for institutional use, such as within universities or corporate environments, open-source models offer a viable solution to mitigate privacy concerns, as they can be hosted on-premises, ensuring that sensitive data does not leave the organization.

Limitations encountered

We encountered several limitations while testing the creativity of open-source models that are listed below:

1. **Long Inference Time:** One of the primary challenges was the considerably longer inference time. These models often required 2-5 times more processing time compared to GPT-3.5 and GPT-4. For example, GPT-4 would take approximately 20 seconds to generate 30 ideas while it takes approximately 1min for Llama2 and 3min for Vicuna (run on local server). This increased latency can be a significant drawback in real-time applications where quick responses are essential.

2. **Prompt Compliance Issues:** The open-source models did not consistently follow the prompts designed for GPT-3.5 and GPT-4. This inconsistency resulted in outputs that varied widely in structure and content, complicating the evaluation pipeline. Effective use of these models would necessitate developing adaptive post-processing functions tailored to each model's unique behavior.

3. **Need for Custom Prompts:** From our preliminary tests, we concluded that each open-source model would require specific prompts to achieve optimal performance. Unlike the more generalizable prompts used with GPT-3.5 and GPT-4, we found that these models needed highly customized prompts, which added complexity to their deployment and integration into our workflows.

In summary, while open-source LLMs like Mistral 7B, Vicuna 13B, and Llama2 7B offer potential benefits in terms of cost and privacy, they also present significant challenges. The increased inference time and need for model-specific prompts and post-processing functions pose hurdles that need to be addressed for these models to be viable alternatives in practice.

3.3 Reproducibility and Consistency

3.3.1 OCSAI

First we tested the reproducibility of OCSAI results, namely that it correlates up to 0.81 with human raters as mentioned in paper of Organisciak et al. 2023 and table 2.3. To this end, we call the OCSAI API to automatically rate the originality of human-generated ideas. We did it with only 10% of the dataset (after shuffling all the samples to have the four different objects) due to time constraints. In fact, OCSAI rates 1000 ideas in approximately 30 min, so it would have taken 3h to score all ideas. As shown in Fig 3.5, the correlation between originality scores of humans and OCSAI is 0.61, which is satisfying and not too far from the different results mentioned in the paper of the authors.

Secondly we tested the consistency of OCSAI by generating 10 ideas for the same object using GPT-3.5. Each idea was then rated 10 times separately by OCSAI. The results demonstrated that OCSAI consistently rated each idea with the exact same score across all 10 evaluations. This consistency in OCSAI's scoring indicates that the model reliably assesses creativity, providing stable and reproducible ratings for the same inputs. Such reliability is crucial for our application where OCSAI was directly integrated into the evaluaton pipeline of the creative ideas.

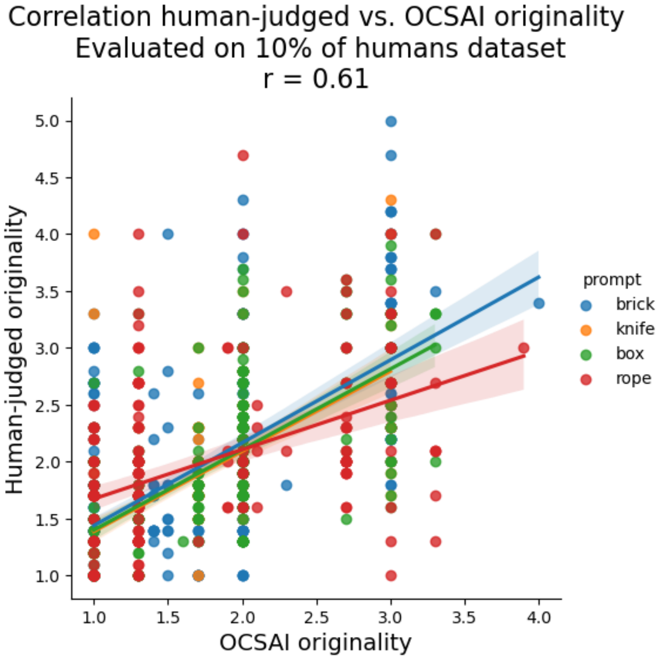


FIGURE 3.5
Correlation human-judged vs. OCSAI originality evaluated on 10% of humans dataset

3.3.2 LLMs

In this part, we aim to assess the consistency and reliability of large language models (LLMs), specifically GPT-3.5 and GPT-4, in generating creative ideas. By evaluating the reproducibility and stability of their outputs across multiple runs, we seek to understand how reliably these models can produce creative content and how representative our findings are.

To this end, we prompted 10 consecutive times GPT-3.5 and GPT-4 (with our initial prompt P1) to generate 100 alternatives uses for the object *brick*. Fig 3.6 below shows the consistency of the GPT models' outputs based on their originality and elaboration. Note that we did not end up with exactly 100 ideas per run because the GPT models often did not strictly respect the number of uses requested explicitly in the prompt. Interestingly, we see that GPT-4 tend to respect this instruction slightly better (92 ideas per run generated instead of 100) than GPT-3.5.

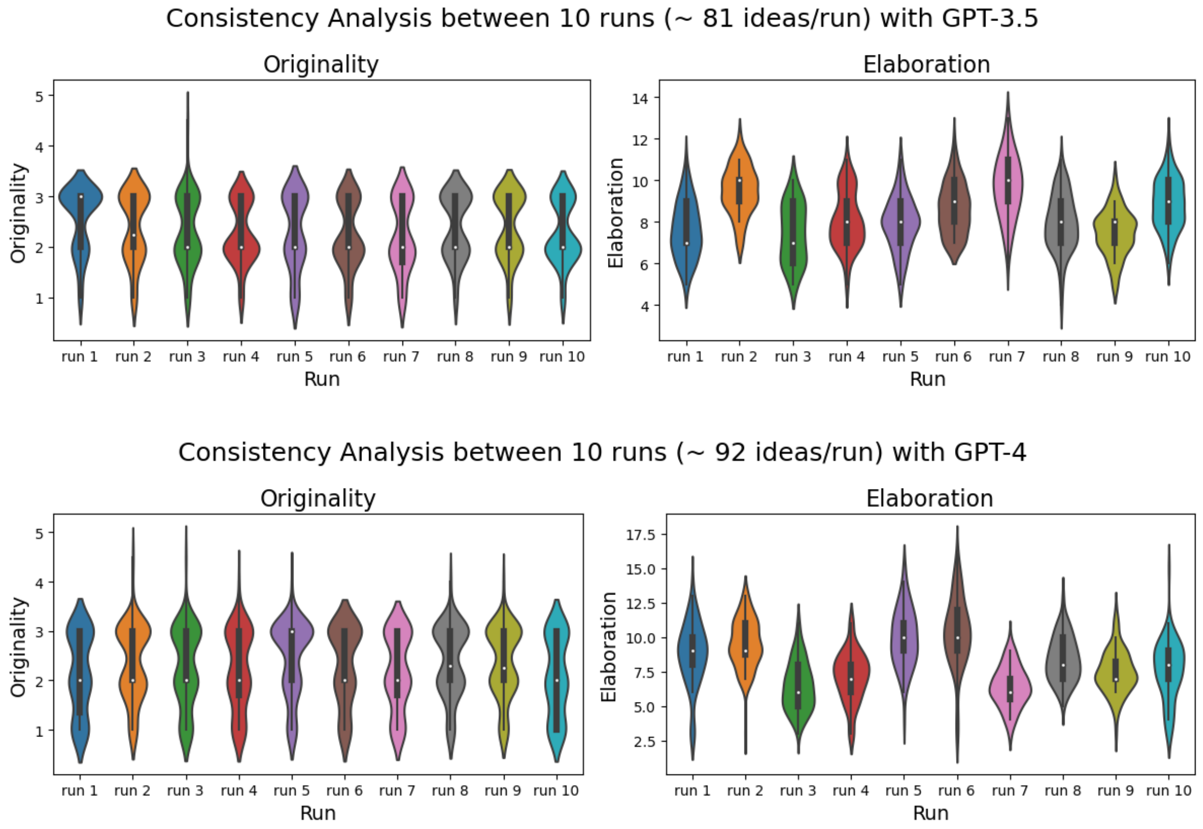
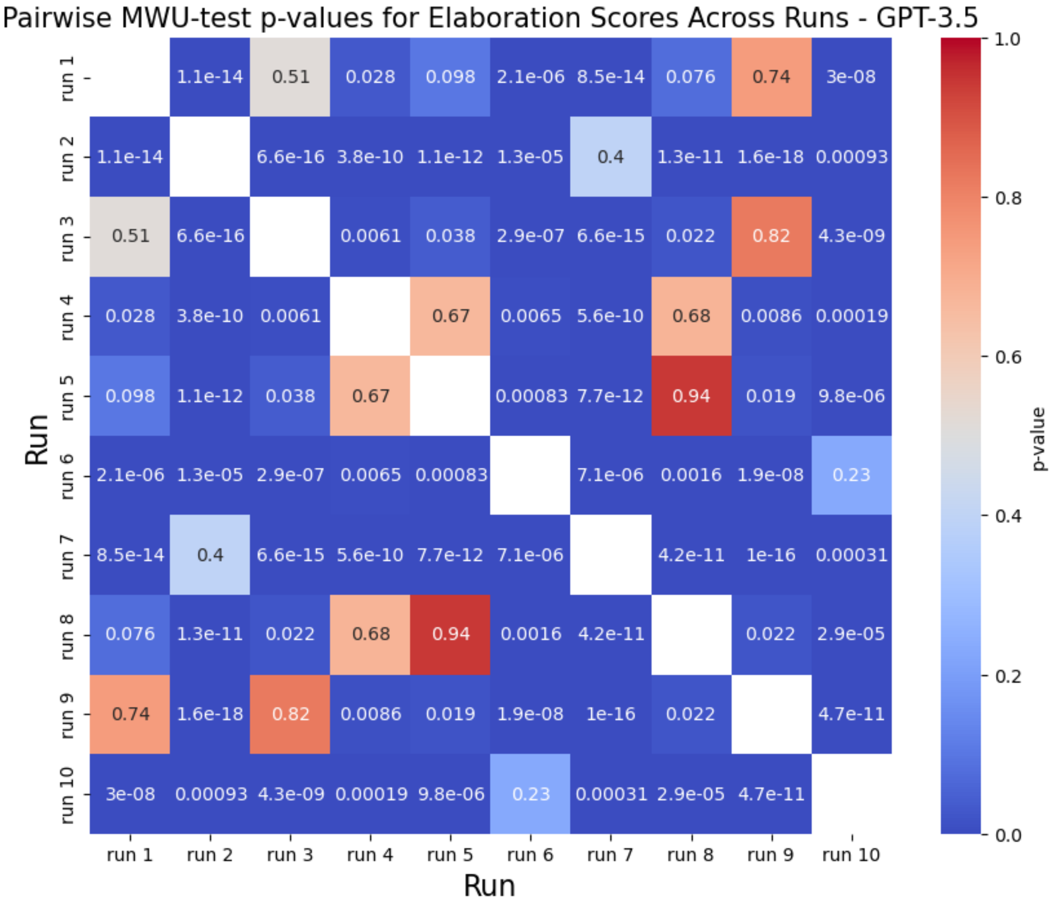


FIGURE 3.6
Consistency Analysis of LLMs outputs

In our analysis, we used the Mann-Whitney U (MWU) test to assess the consistency between the outputs of GPT models across ten different runs by comparing the elaboration distribution. We used the elaboration metric because it does not depend on OCSAI like the originality and we suppose that it can correctly assess how different generations of ideas are different or similar in their structure. The MWU test is a non-parametric test used to determine whether two independent groups have different distributions of a continuous variable. We chose this test because it does not assume normality of the data nor equal variances (two conditions that the elaboration distributions of the 10 runs did not satisfy).

Analysing Fig 3.7, we observed that most of the p-values were below the significance threshold of 0.05, indicating statistically significant differences in elaboration scores between all pairs of runs. This finding suggests that the outputs of GPT-3.5 varied significantly across different runs in terms of elaboration, indicating inconsistency in the model's performance across runs. Therefore, we can conclude that there is considerable variability in outputs across different runs, highlighting the importance of assessing model consistency and robustness in natural language generation tasks.

We performed the same analysis for GPT-4 (see Fig. 5.6 in Appendix) and had the same conclusion.



Run	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
run 1		1.1e-14	0.51	0.028	0.098	2.1e-06	8.5e-14	0.076	0.74	3e-08
run 2	1.1e-14		6.6e-16	3.8e-10	1.1e-12	1.3e-05	0.4	1.3e-11	1.6e-18	0.00093
run 3	0.51	6.6e-16		0.0061	0.038	2.9e-07	6.6e-15	0.022	0.82	4.3e-09
run 4	0.028	3.8e-10	0.0061		0.67	0.0065	5.6e-10	0.68	0.0086	0.00019
run 5	0.098	1.1e-12	0.038	0.67		0.00083	7.7e-12	0.94	0.019	9.8e-06
run 6	2.1e-06	1.3e-05	2.9e-07	0.0065	0.00083		7.1e-06	0.0016	1.9e-08	0.23
run 7	8.5e-14	0.4	6.6e-15	5.6e-10	7.7e-12	7.1e-06		4.2e-11	1e-16	0.00031
run 8	0.076	1.3e-11	0.022	0.68	0.94	0.0016	4.2e-11		0.022	2.9e-05
run 9	0.74	1.6e-18	0.82	0.0086	0.019	1.9e-08	1e-16	0.022		4.7e-11
run 10	3e-08	0.00093	4.3e-09	0.00019	9.8e-06	0.23	0.00031	2.9e-05	4.7e-11	

FIGURE 3.7
Consistency Analysis of GPT-3.5 outputs with Mann-Whitney U test on Elaboration

3.4 OTHER EXPLORATION PATHWAYS

3.4.1 ON A BENCHMARK LEVEL

On a benchmark level, several avenues of research could be explored for each dimension.

For the measurement of flexibility in our benchmarking framework, one approach could be to explore different topic modeling strategies. Currently, our project may rely on a specific topic modeling algorithm (LDA), but experimenting with alternatives such as Non-negative Matrix Factorization (NMF) or even newer deep learning-based models could yield richer insights into the topical diversity of generated ideas. Additionally, varying the number of topics used in the analysis and redesigning the flexibility score could further enhance our understanding of how flexible the ideas are. Implementing multi-topic assignment instead of assigning a unique topic to each idea would also provide a more nuanced view of flexibility, capturing ideas that cover multiple topics without having to deal with overlaps.

Our current method for evaluating elaboration, which involves removing stop words, could be expanded to incorporate more advanced techniques. For instance, leveraging natural language processing (NLP) tools to analyze sentence complexity, syntactic variety, and the use of descriptive language could offer a deeper assessment of how elaborative the generated ideas are. Implementing these advanced techniques would align our evaluation closer with the qualitative aspects that human evaluators consider when assessing elaboration.

In terms of dissimilarity, exploring alternative models for calculating cosine similarity could provide more robust assessments. While sentence embeddings have been useful, other approaches using different pretrained language models with different sizes could improve the accuracy and reliability of dissimilarity measurements. Additionally, exploring alternative methodologies beyond sentence embedding, such as graph-based representations of ideas or clustering approaches, could offer new perspectives on how distinct the generated ideas are from one another.

3.4.2 ON THE LLMS LEVEL

A comprehensive and rigorous comparison between the outputs of the GPT models used is essential for understanding their capabilities and limitations. An ablation study, which systematically removes or alters components of the prompts, could reveal the impact of specific features on the generated alternative uses.

Investigating the role of examples choice in few-shot prompting is another critical pathway. Our preliminary findings suggest that the examples used in the prompts can significantly influence the originality and quality of generated ideas. Future work could systematically vary the examples to identify optimal strategies for enhancing creativity. Moreover, directly introducing the models to our creativity dimensions (originality, elaboration, flexibility, dissimilarity) and explicitly prompting them to maximize these dimensions could lead to more targeted and effective ideas generation.

Investigating various prompting strategies and, optionally, post-processing functions to maintain consistent outputs from open-source models would significantly enhance this research project. Additionally, examining other factors that influence prompts could prove useful, such as the impact of the temperature parameter on the consistency of outputs from LLMs, or exploring alternative frameworks of prompts.

A comparative analysis could also be conducted between the different GPT models and open-source models like Mistral 7B, Vicuna 13B, and Llama2 7B. This comparison could highlight the strengths and weaknesses of each model type, informing decisions on which models are best suited for specific applications in educational settings and beyond.

CHAPTER 4

CONCLUSION

This study sets out to explore how LLMs could be leveraged to generate creative ideas, particularly focusing on the Alternative Uses Task (AUT). We developed a comprehensive methodology to evaluate the creativity of outputs from large language models (LLMs) along multiple dimensions. By comparing human-generated ideas with those produced by GPT-3.5, GPT-4, and various open-source models, we aimed to understand the nuances of LLM behavior in creative tasks.

Our findings reveal that while LLMs can generate creative ideas, their outputs are heavily influenced by the specific prompts used. Different prompting strategies, including few-shot examples, can significantly impact the originality, elaboration, flexibility and diversity of the generated ideas. We observed that GPT models, particularly with tailored prompts, can achieve a higher level of creativity than initially expected. However, the performance of open-source models like Mistral 7B, Vicuna 13B, and Llama2 7B was hindered by longer inference times and the need for custom prompts, highlighting areas for further optimization.

From our study, several key recommendations emerge for generating creative ideas using the AUT. Effective prompting, such as using few-shot examples with high originality, tends to yield better results. Moreover, ensuring prompts are clear and specific to the task can enhance the quality of the generated ideas. While LLMs can produce creative outputs, they still struggle with consistency and diversity, especially when generating a large number of ideas at once.

Looking ahead, there is significant potential to transfer these findings to more concrete and impactful problems. By refining our understanding of how to prompt LLMs effectively and designing comprehensive evaluation metrics, we could harness their creative potential for broader applications. We believe that the insights discovered in this study could extend to various open-ended problems, suggesting that with further advancements, LLMs could become powerful tools for innovative problem-solving in diverse fields.

CHAPTER 5

APPENDIX

5.1 DATASET

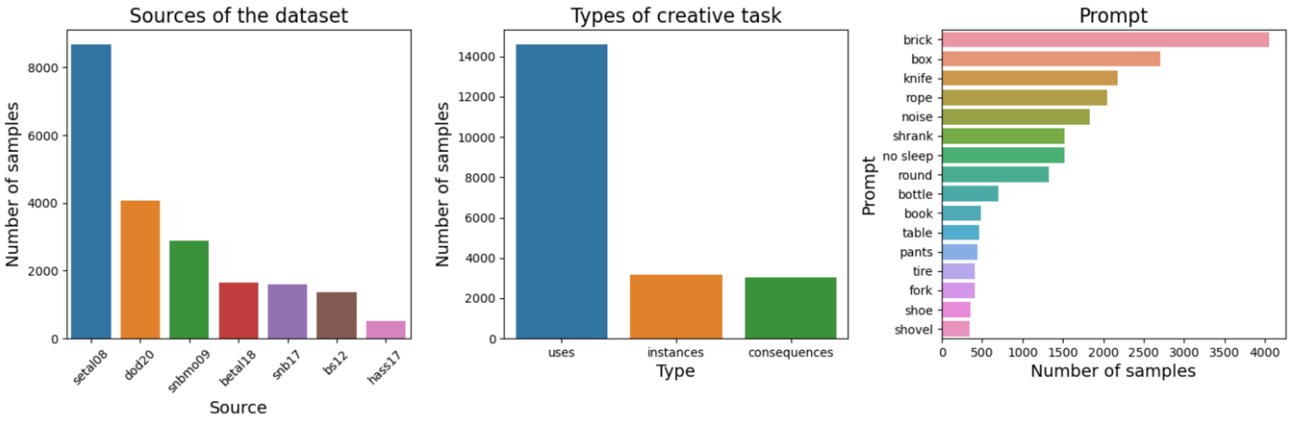


FIGURE 5.1
EDA of original dataset

5.2 FLEXIBILITY

In more details, the coherence metric follows a four-stage pipeline as described by Röder, Both and Hinneburg 2015:

- **Segmentation:** Splits the set of words into pairs.
- **Probability Estimation:** Estimates the probability of word co-occurrence.
- **Confirmation Measure:** Measures the degree of confirmation between word pairs.
- **Aggregation:** Aggregates individual scores to produce a coherence score.

We specifically used the c_v coherence score.

Overlaps of topics' keywords

The plots below show the overlap of topics' keywords for each object for different configurations of number of topics used for the LDA and different number of keywords per topic.

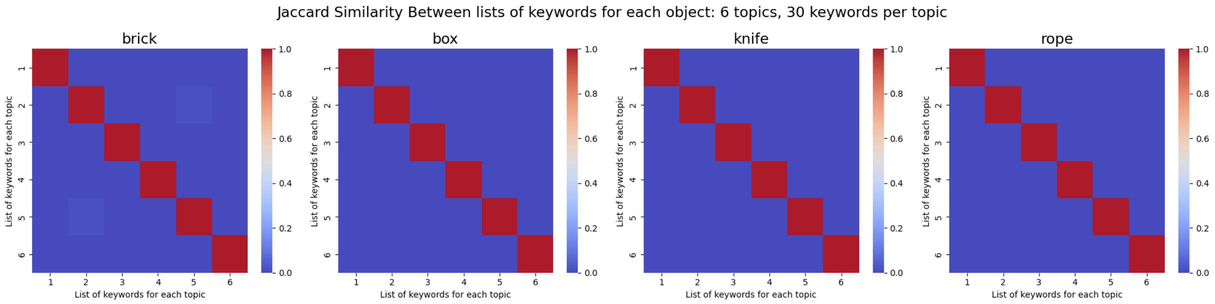


FIGURE 5.2
Overlap of keywords with 6 topics and 30 keywords per topic, one can see a small overlap of keywords between topics 2 and 5 for object brick

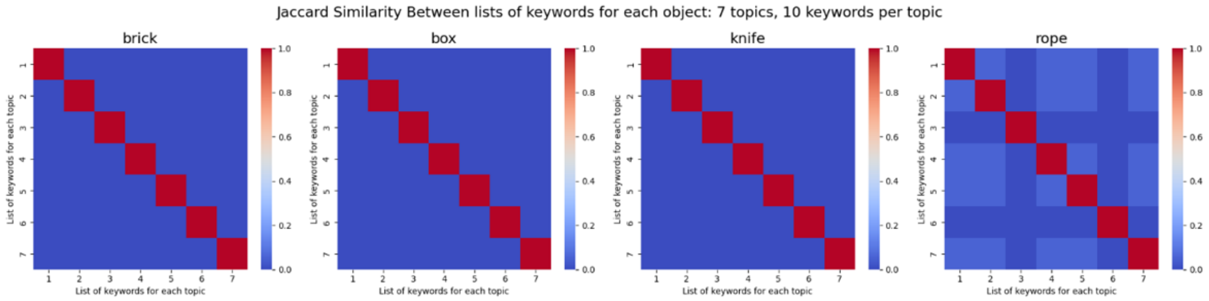


FIGURE 5.3
Overlap of keywords with 7 topics and 10 keywordsper topic, one can see an important overlap of keywords between topics 1, 2, 4, 5 and 7 for object rope

5.3 ELABORATION

Part of Speech Tagging results

Part of speech (POS) tagging is the process of identifying and labeling the words in a sentence with their corresponding part of speech, such as nouns, verbs, adjectives, adverbs, etc. This is typically done using natural language processing (NLP) techniques and algorithms that analyze the grammatical structure and context of each word.

We applied POS tagging to the ideas generated by the Humans as well as the LLMs by using the specific NLTK built-in function[1]. As shown in Fig 5.4, we cannot see a real difference between the Humans and the LLMs on a grammatical structure level when applying POS tagging, which led us to left this option.

[1] https://www.nltk.org/api/nltk.tag.pos_tag.html

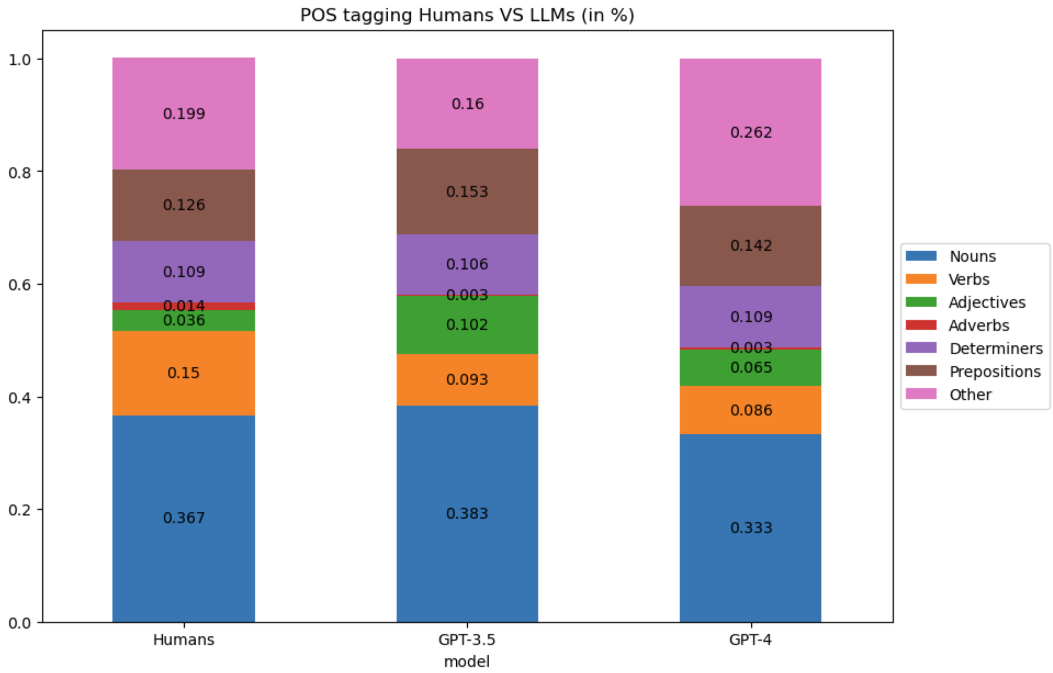


FIGURE 5.4
Part-of-Speech Tagging

5.4 PROMPTS

FRAMEWORK (REMAINING PROMPTS)

Initial prompt without length constraint (P3)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Initial prompt without persona (P4)

You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Initial prompt without persona and context (P5)

Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should be a concise sentence and follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

Initial prompt without creative and length constraints (P7)

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original.
Generate exactly N_responses alternative uses for the object [object_name].
Each alternative use should follow the same format as the examples below:
Sock, Color it and maybe make a snake
Sock, Use it as a puppet

FEW-SHOT EXAMPLES

Example of a prompt with 5 few-shot examples with maximum originality

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly 100 alternative uses for the object [brick].
Each alternative use should be a concise sentence and follow the same format as the examples below:
brick, make it into a superhero called The Wall and have him fight Erosion Man
brick, add a lens to the wholes in a and create some sort of binocular or glasses
brick, use it to create a pattern in paint and then stamp in onto paper
brick, Crush a into a fine sand-like powder and use it as a makeshift water filter
brick, for water displacement in a science experiment

Example of a prompt with 5 few-shot examples randomly selected with scores

You are a very creative, open-minded person who can propose creative, out-of-the-box ideas while staying realistic.
You are meant to assist students in group ideation. They are asked to propose alternative uses for an object, and you should share your ideas of alternative uses to inspire them to explore other possibilities. Your ideas will be especially appreciated if they are original, useful in real life, or both.
Generate exactly 100 alternative uses for the object [brick].
Each alternative use should be a concise sentence and follow the same format as the examples below:
brick, place flowers in the holes of the brick
brick, hot "" instead of potato
brick, support for holding a fishing pole after it has been casted.
brick, Polish the sell for profit.
brick, block a door from closing/opening

Below are the originality scores for each example listed in order. The scores are on a scale from 0 to 5, with 0 being not original and 5 being very original. Use these scores to understand what makes an idea original, but do not include them in your output.
Originality scores: [2.0, 1.8, 3.0, 2.6, 1.3]

5.5 COMPARISON PROMPTS PER OBJECT FOR GPT-3.5

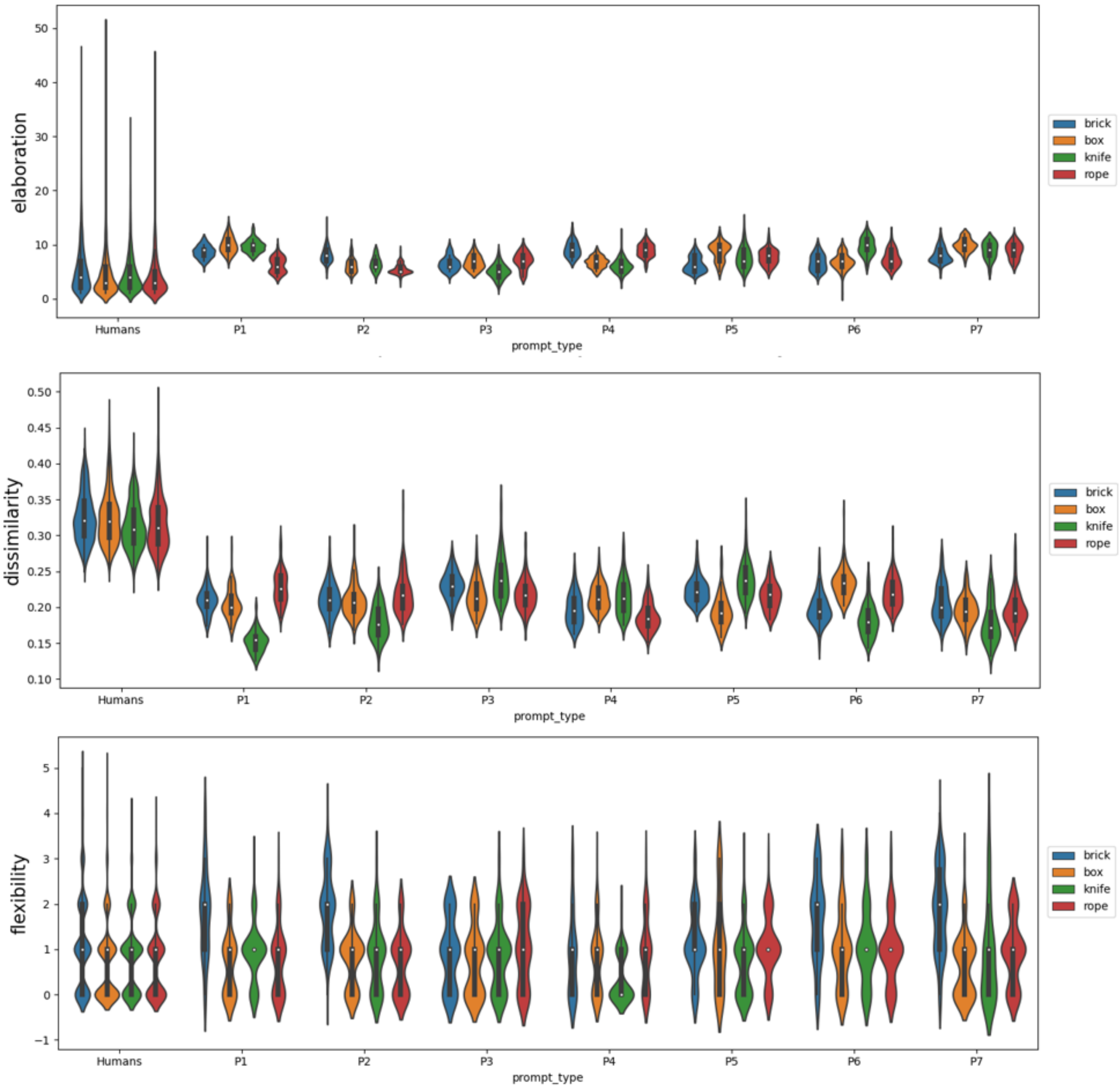
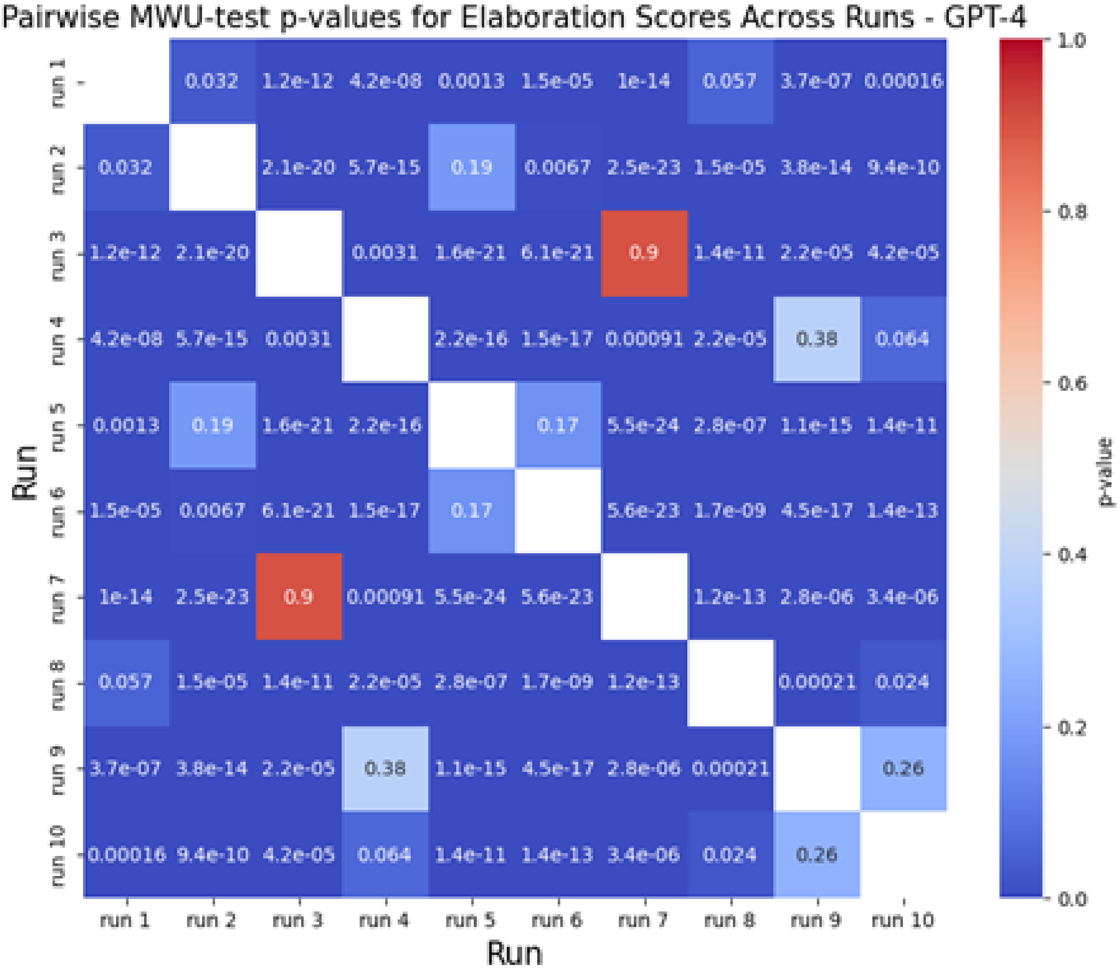


FIGURE 5.5
Comparison prompts per object on elaboration (top), dissimilarity (middle), flexibility (bottom) for GPT-3.5

5.6 CONSISTENCY



	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
run 1		0.032	1.2e-12	4.2e-08	0.0013	1.5e-05	1e-14	0.057	3.7e-07	0.00016
run 2	0.032		2.1e-20	5.7e-15	0.19	0.0067	2.5e-23	1.5e-05	3.8e-14	9.4e-10
run 3	1.2e-12	2.1e-20		0.0031	1.6e-21	6.1e-21	0.9	1.4e-11	2.2e-05	4.2e-05
run 4	4.2e-08	5.7e-15	0.0031		2.2e-16	1.5e-17	0.00091	2.2e-05	0.38	0.064
run 5	0.0013	0.19	1.6e-21	2.2e-16		0.17	5.5e-24	2.8e-07	1.1e-15	1.4e-11
run 6	1.5e-05	0.0067	6.1e-21	1.5e-17	0.17		5.6e-23	1.7e-09	4.5e-17	1.4e-13
run 7	1e-14	2.5e-23	0.9	0.00091	5.5e-24	5.6e-23		1.2e-13	2.8e-06	3.4e-06
run 8	0.057	1.5e-05	1.4e-11	2.2e-05	2.8e-07	1.7e-09	1.2e-13		0.00021	0.024
run 9	3.7e-07	3.8e-14	2.2e-05	0.38	1.1e-15	4.5e-17	2.8e-06	0.00021		0.26
run 10	0.00016	9.4e-10	4.2e-05	0.064	1.4e-11	1.4e-13	3.4e-06	0.024	0.26	

FIGURE 5.6
Consistency Analysis of GPT-4 outputs with Mann-Whitney U test

Bibliography

Guilford, Joy Paul (1950). ‘Fundamental statistics in psychology and education’. In.

Runco, Mark A and Selcuk Acar (2012). ‘Divergent thinking as an indicator of creative potential’. In: *Creativity research journal* 24.1, pp. 66–75.

Stevenson, Claire et al. (2022). ‘Putting GPT-3’s creativity to the (alternative uses) test’. In: *arXiv preprint arXiv:2206.08932*.

Beaty, Roger E and Dan R Johnson (2021). ‘Automating creativity assessment with SemDis: An open platform for computing semantic distance’. In: *Behavior research methods* 53.2, pp. 757–780.

Organisciak, Peter et al. (2023). ‘Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models’. In: *Thinking Skills and Creativity* 49, p. 101356.

Runco, Mark A and Garrett J Jaeger (2012). ‘The standard definition of creativity’. In: *Creativity research journal* 24.1, pp. 92–96.

Nijstad, Bernard A et al. (2010). ‘The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence’. In: *European review of social psychology* 21.1, pp. 34–77.

Blei, David M, Andrew Y Ng and Michael I Jordan (2003). ‘Latent dirichlet allocation’. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Zhang, Tianyi et al. (2019). ‘Bertscore: Evaluating text generation with bert’. In: *arXiv preprint arXiv:1904.09675*.

Brown, Tom et al. (2020). ‘Language models are few-shot learners’. In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Röder, Michael, Andreas Both and Alexander Hinneburg (2015). ‘Exploring the space of topic coherence measures’. In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.