Automating Alternative Generation in Decision-Making

Yevhen Kostiuk, Clara Seyfried, Chris Reed
Centre for Argument Technology (ARG-tech)
University of Dundee, United Kingdom
kosteugeneo@gmail.com {c.a.reed, c.seyfried}@dundee.ac.uk

Abstract

In decision making, generating alternative solutions is crucial for solving a problem. However, cognitive biases can impede this process by constraining individual decision makers' creativity. To address this issue, we introduce a new task for automatically generating alternatives, inspired by the process of human "brainstorming". We define alternative options based on atomic action components and present a dataset of 106 annotated Reddit *r/Advice* posts containing unique alternative options extracted from users' replies. We also introduce new metrics to assess the quality of generated components, including distinctiveness, creativity, upvote-weighted, crowd intersection, and final commit intersection scores. As a baseline, we evaluated the large language models (LLMs) LLaMa3:8b, LLaMa3.1:8b, and Gemma 2:9b on the alternative component generation task. On the one hand, models demonstrated high creativity (ability to generate options beyond what Reddit users suggested) and performed well at proposing distinct alternatives. A subset of generated components was manually evaluated and found overall useful. This indicates that LLMs might be used to extend lists of alternative options, helping decision makers consider a problem from different perspectives. On the other hand, LLMs' outputs often failed to align with human suggestions, implying that they still tend to miss important components.

The code and annotation guidelines can be found in the project's GitHub repository[1]. The dataset is available via the link[2].

1 Introduction

Decision-making is the process of choosing any course of action that aims to solve a problem in the best possible way. Every aspect of human life involves decision making to some degree, from selecting what to wear based on the weather to deliberating how to resolve a large-scale conflict. Yet, often human creativity is limited and constrained by biases when it comes to imagining different alternative actions leading to the best possible outcome.

Although different theories disagree on the structure of the decision making process (Morelli et al., 2022), most theoretical frameworks consider decision making to be the process of selecting a preferred option from a set of alternative options, frequently without specifying from where this set of alternatives originates. It has been highlighted in multiple studies that this process is a very important yet challenging step (Hämäläinen et al., 2024; Fisher et al., 1983).

As it relies on memory retrieval of relevant information (Johnson et al., 1991), the process of identifying possible actions poses human challenges. For example, information may not be organized in a coherent structure useful for the problem (Jungermann et al., 1983), interference from previous knowledge can make it more difficult to restructure information to see problems from different perspectives (Heuer, 1999), and the use of heuristics based on prototypical problems may introduce further human biases such as a tendency to learn towards highly representative options, and struggling to retrieve or creatively generate high-quality solutions (Gigerenzer and Gaissmaier, 2011; Gettys et al., 1987).

There are multiple theoretical frameworks developed to help decision makers overcome these challenges (Keeney, 1992; Pitz et al., 1980). "Brainstorming" is one of them (Al-Samarraie and Hurmuzan, 2018; Hicks, 1991). This method relies on aggregated judgment from multiple people, trying to utilize their cognitive efforts and mitigate their individual biases by providing different opinions and perspectives on the problem at hand. Overcom-

[1] https://github.com/arg-tech/decisionmaking_alternative_components

[2] https://huggingface.co/datasets/arg-tech/decisionmaking_alternative_components

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1–15
November 4-9, 2025

ing humans' limited cognitive flexibility, "brainstorming" is also particularly well-suited for computational automation (John, 2016).

Developing an automatic algorithm can help overcome human biases by generating wider set of possible alternatives. Such algorithms could be used in any process or task requiring decision making, such as operational planning, conflict resolution, and so on.

In this paper, we introduce a novel task in which the algorithm needs to generate lists of possible atomic options for solving human decision making problems. As a baseline, we evaluate different large language models (LLMs) in few- and zero-shot settings. To evaluate the performance of the models, a new dataset based on Reddit comments was manually labeled, approximating the "brainstorm" technique by incorporating advice replies from multiple users per question. The following contributions are made:

- We propose a new definition of alternative options based on atomic units of action (*components*), and introduce the **Component Generation (CG)** task.
- We present a new dataset for the CG task based on the Reddit Advice subreddit[3] (r/Advice). Specifically, we filtered posts requesting advice based on predefined criteria. Then we extracted, labeled, and summarized proposed potential solutions from the comments for the filtered posts, and marked if the comment author considered them to be competing (mutually exclusive). Additionally, we identified comments to which the post author responded, extracting both atomic actions and whether the author did or committed to doing the proposed action, providing an indicator of which alternatives were perceived to be particularly helpful.
- We introduce novel metrics for evaluating alternative generation: **distinctiveness**, **creativity**, **upvote-weighted intersection**, **crowd intersection**, and **final commit intersection** scores. The matching algorithm was manually validated.
- We evaluate different LLMs for alternative generation using both zero-shot and few-shot approaches. For few-shot approaches, we used 5 and 10 examples, averaging results across three trials with different sets of examples. We conducted experiments with LLaMa3:8b (Dubey et al., 2024), LLaMa3.1:8b (Dubey et al., 2024), and Gemma 2:9b (Team, 2024).

[3] https://www.reddit.com/r/Advice/

2 Related Work

No datasets or evaluation metrics were available for the task at the time of writing. Decision-making has been discussed in academic literature from various disciplines, yet there is a notable lack of materials when it comes the generation of alternatives from a computational perspective.

Early studies (Arbel and Tong, 1982; Ozernoy, 1985; Alexander, 1979) were focusing on the benefits and influence of alternative generation on the overall decision process. A comprehensive survey and overview of different techniques (Keller and Ho, 1988) discussed in detail various different methods that a person could utilize to achieve the best possible set of alternative actions in any given scenario.

"Mean-value" approaches (Keeney, 1992; Keeney et al., 1994; León, 1999) encourage decision makers to estimate the relative importance of alternatives ("values") as well as the means to achieve them ("means"). In this framework, the alternative choices are considered to be given. Similarly, in one of the most well-known models of decision making so far, Simon (1955) introduced a "design" concept prior to the "choice".

MGA (Modeling for Generating Alternatives), a theoretical framework for alternative generation, has been presented and discussed in various studies (Brill Jr et al., 1982; Chang et al., 1983; Simon, 1955; DeCarolis, 2011; DeCarolis et al., 2016). The proposed algorithm formalizes a decision making process. The method includes multi-objective optimization algorithms to explore the neighborhood of a possible solution in order to find the most optimal solution. It requires a distance function initialization that measures the differences between the solutions, as well as a strict definition of the importance of the model's objectives and constraints. Recently, Colorni and Tsoukiàs (2020) formulated a general framework for formalizing alternatives, stressing how under-researched this area has been.

The first attempts at systems practically aiding the decision making process were introduced in the late 90s (Leal and Pearl, 1977; Pearl et al., 1982;

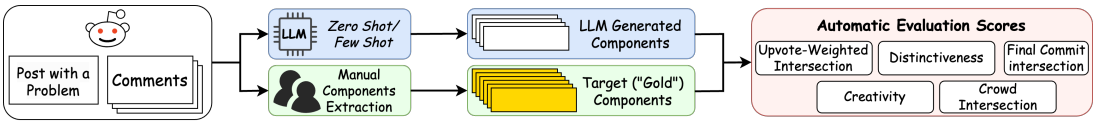

Figure 1: Pipeline Overview of the Framework.

Leal et al., 1978; Steeb and Johnston, 1981; Keller and Ho, 1988). Recent studies address further theoretical subtleties, carefully structuring various techniques and methods that can aid decision makers (Hämäläinen et al., 2024). It is furthermore worth noting that research incorporating the use of artificial intelligence to assist idea generation generally (Shaer et al., 2024) is also thematically adjacent to our research, though it lacks the distinction between different competing alternatives which is so crucial for describing decision making.

Lastly, the proposed task also shares some similarity with question answering tasks (QA), for which there are different datasets available (quo, 2017; Lovenia et al., 2024; Kwiatkowski et al., 2019; Rajpurkar et al., 2016; Reddy et al., 2019; Joshi et al., 2017; Yang et al., 2018; Talmor et al., 2019). However, QA tasks have a correct answer, whereas in the case of alternatives this is not necessarily clearly evident.

3 Definitions and Tasks

In this work, an *alternative option* (AO) refers to an action or a set of actions that could be taken in the context of a given problem. An action refers to the specific steps taken to implement the chosen alternative, aiming to resolve the problem and achieve the desired outcome. In the context of this work, a problem refers to some type of situation that requires a response and involves a choice between different AOs to achieve or solve it, based on a definition by Beachboard and Aytes (2013). In decision making, solutions are considered side-by-side and can therefore be termed "alternatives". Each AO consists of smaller units of action that we call *components*, which are atomic, i.e., cannot be broken down further. More formally, components are the smallest actions that can be taken in solving the problem. They may include conditional elements, a certain order of components, concretize specific actors who should participate or perform the component, etc. Components are characterized by the order or actions, semantic content, conditional parts, and the entities and participants included in the components. Components c_1 and c_2 are considered identical if they fit the following **component matching rules (CMRs)**, i.e., if c_1 and c_2 preserve the order of actions and overall semantics, maintain the same conditional parts, and refer to the same entities, participants, people, etc. Consider the following example:

Problem:
An office worker A accidentally took a cookie from a bowl in the office kitchen, assuming that it was a shared bowl. However, it turns out that the contents of the bowl were the private lunch of colleague B, who still has not noticed that one of the cookies is missing. What do you recommend worker A should do in this situation ?

In this scenario, the components could be as follows:

1. Do not tell B that you took the cookie.
2. Tell B that you took the cookie.
3. Buy B a whole new pack of cookies.
4. If you value B's friendship, tell the truth.
5. Take more cookies.
6. Tell B that you took the cookie, then see B's reaction: if B is angry, buy B a lunch.
7. Buy B a lunch.
8. Tell B and HR that you took the cookie.

In the example, there are multiple suggested component actions that A can take. For example, Components 2 and 4 are different, as there is an additional conditional part to Component 4 that is not present in 2. Components 2 and 8 include the same action, but different participants (Component 2: B; Component 8: B and HR). Component 6 contains an order of actions. Some of the components can be done together (e.g., Components 1 and 5). If the components are mutually exclusive (e.g. Components 1 and 2), they are referred to as *competing components*. Each AO is a set of non-competing components, i.e., all of its actions can be performed without a conflict. *Competing*

alternatives are sets of AOs whose components are competing with each other. This is similar to the concept of competing hypotheses (Heuer, 1999).

For instance, because of the competing Components 1 and 2, alternative options $AO_1 = \{1, 3, 7\}$ and $AO_2 = \{2, 3, 7\}$ are competing alternatives, despite having overlapping components. A subset of the AO is an AO itself.

The **Component Generation (CG)** task involves generating as many relevant components as possible to solve the problem, based on its title and description. This is framed as a text generation task.

In this paper, we focus on the CG task and leave other potential tasks for future work.

4 Dataset

4.1 Annotation

In this work, we introduce a novel dataset which was specifically created to fit the task. For this, we manually labeled the Reddit Corpus dataset from the ConvoKit (Chang et al., 2020). From this dataset, all of the *r/Advice* subreddit posts and comments were gathered for annotation. Two annotators were recruited for the task of labeling alternatives: a PostDoc researcher (Annotator a) and a PhD student (Annotator b) from the Centre for Argument Technology at the University of Dundee (UK).

The annotators were presented with the following task. In the Doccano (Nakayama et al., 2018) annotation system they were presented with a Reddit post (title and content) from the dataset and a list of comments: the 50 most upvoted comments, 25 random comments and pairs of comments (original post's author reply and the comment that this reply was addressed to). The annotators were asked to read the post and determine if in the post the problem was stated clearly and if the author was asking for help in a search of possible actions or options to take. Posts that did not meet these criteria, asked for medical or legal advice, required very specific domain knowledge, included moral dilemmas, included images, or were too broad (e.g. open-ended questions such as "What should I name my cat?") were disregarded. The rules for post exclusion were created empirically from the initial labeling by the authors of the paper. If the post fitted the criteria, the annotator checked all comments and determined if users proposed a solution in any of the comments. For each comment in which a potential solution could be found, the solution was split into its respective components. Each component was summarized to retain as much information as possible while keeping it short (e.g. this involved reconstructing pronouns and anaphora, removing hate speech etc.). The annotators were instructed **not** to consider the quality of the components, but to remove clearly sarcastic, joke, and offensive propositions. After summarizing, the annotators compared each component with the list of previously retrieved components for this post. If the component had already been proposed, the exact phrasing of the already existing component was taken from the list of alternatives and assigned to the new comment. If the component had not been proposed yet, it was assigned to the comment and added to the list. Additionally, if the author of a comment implied that multiple components of their alternatives were competing, the annotators noted the mutual exclusivity of these components by marking each competing solution with an alternative number for the comment. For example, if a comment proposed that the author of the original post should choose one of two options (e.g. *Do this ... or do that ...*), the solution was marked with the numbers 1 and 2 with respect to the mentioned components. If there was only one proposed solution, it was marked with -1. A new list of alternatives was created for each new post. The annotation guidelines can be found in the project's GitHub repository.

On Reddit, authors can edit their original post with an update after it was posted. For the annotation, we removed updates from the posts, as they usually were not available to the commentators when they proposed their solutions. The text of the update part was annotated separately, as the author response was relevant for later analyses of which alternatives the author committed to.

4.2 Inter-Annotator Agreement

To measure inter-annotator agreement, we used two strategies: comparing the number of extracted components per post and the components' semantic intersection with the respect to the component matching rules (CMRs).

15 posts with 148 comments were included in the annotation data for both of the annotators (as an overlap). The total number of comments which contained at least one component was 79 for Annotator a and 120 for Annotator b. Across all posts, the total number of unique components was 49 for

Annotator a and 51 for Annotator b. The Cohen's kappa score (Kohen, 1960) for the count of components extracted per comment was 0.614, indicating substantial agreement (Landis and Koch, 1977).

To answer the question if annotators extracted the same components, we used the following approach. For each of 15 posts, Annotator a was presented with two lists of components: unique components that extracted by Annotator b and unique components extracted by Annotator a. For each component from the a's list, the annotator marked which component (if any) from b's list it might correspond to based on the CMRs. To calculate an agreement, we used the following formula:

$$\text{CMR}_{\text{agr}} = \frac{\sum_p (U_a^p + U_b^p)/(T_a^p + T_b^p)}{N},$$

where p is the post, U_a^p is the number of components from Annotator a's unique list of components that did not appear in Annotator b's list for the post p. U_b^p is the number of components from Annotator b's unique list of components that did not appear in Annotator a's list for the post p. T_a^p and T_b^p refer to the total amount of unique components extracted by annotators a and b respectively for the post p. N is the total number of posts. Taken together, this score indicates the average amount of components which are extracted by only one of both annotators, with respect to the total number of components. We divide by $T_a^p + T_b^p$ to ensure the fairness of the score.

The lower CMR_{agr}, the higher the annotators' agreement. Our obtained score was 0.36, indicating a reasonable agreement between the similarity of the extracted components.

4.3 Dataset Statistics

After filtering out posts that did not fit the criteria, the total number of unique posts is 106. The total number of unique authors is 101, with 5 posts attached to deleted accounts. The total number of considered comments is 3,828. Among them, the total number of comments that the author of the post replied to is 1,413 in 97 posts. The number of comments that did not propose any solution is 1,999. The average number of solutions per post is 14.03, with the maximum of 44 and minimum of 4. The average number of competing alternatives per post is 1.04 with the maximum of 9. The total number of posts where the author appeared to commit to take some of the actions is 75 (70.7% of all unique posts), with 197 unique components and 240 comments in total (sometimes, the author replied to multiple comments with the same commitment).

	Total	**Mean**	**Unique**
Title	1,803	17.00	687
Post body	41,907	395.34	4,784
Post update	3,704	34.94	1,077
Comment	244,657	68.55	12,421
Component	37,284	8.05	2,949

Table 1: Number of words statistics. The *Post body* value was calculated after the removal of update.

The total number of words can be found in Table 1. To determine the number of words, we used the NLTK (Bird et al., 2009) framework.

5 Evaluation

In the context of the component generation task, we propose the following metrics: **distinctiveness**, **creativity**, **upvote-weighted intersection**, **crowd intersection**, and **final commit intersection** scores.

All the proposed metrics require an algorithm that determines whether the two components are identical. Recall that the components c_1 and c_2 are considered identical if they fit the **component matching rules (CMRs)**: c_1 and c_2 preserve the order of actions and overall semantics, have the same conditional parts, and refer to the same entities, participants, people, etc. The calculation is based on the *components matching algorithm*, which is detailed in the following subsection.

5.1 Components Matching Algorithm

To determine whether the pair of components are identical with respect to the defined CMRs, we developed an LLM-based ensemble method, utilizing an ensemble of LLaMa3:8b (Dubey et al., 2024) and Mistral:7b (Jiang et al., 2023) models. The component matching algorithm architecture is presented in Figure 2.

The component matching algorithm works by providing LLaMa3 with a prompt containing CMRs and four examples covering all the rules. LLaMa3 predicts if any CMR is violated. If it fails to output "MATCH"/"NOT MATCH" answer, the input is passed to Mistral; if Mistral also fails, we assume the CMRs are violated.

We evaluated the component matching algorithm

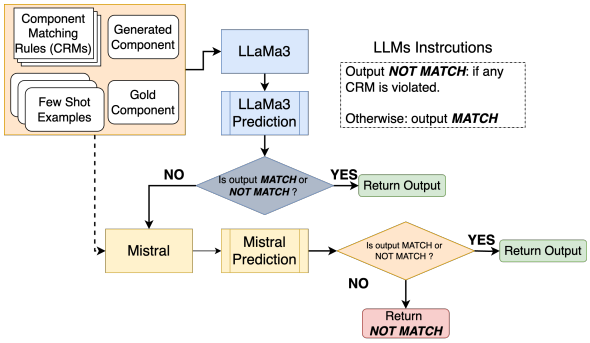

Figure 2: Components Matching Algorithm.

with a Mistral:7b generated dataset of components, which was manually reviewed and filtered. The evaluation and metrics of the component matching algorithm can be found in the Appendix B.1. The limitation of such an approach is that the dataset contains Mistral's biases and could lack variety.

After generating, we manually evaluated a sample of pairs of generated component and gold components. The accuracy of the algorithm was 0.92 and weighted F1 was 0.93. The results are presented in the Appendix B.2. The evaluation of the component matching algorithm showed a lower F1 score of 0.86 on manually labeled data compared to 0.91 on the generated dataset. This difference suggests that the automatically generated dataset may not fully capture the diversity of real matching pairs. It also highlights potential limitations in the algorithm's generalization, warranting further investigation in future work.

5.2 Metrics

Notations Let P_{rs} be a list of all predicted components (repetition possible). with $|P_{rs}|$ referring to the number of elements in this list. P_s is a set of predicted *unique, non-repetitive* components (based on the matching algorithm and CMRs) from the model for the post p. $R_s = \{c_i | i = 1, \ldots, N\}$ is a set of extracted components from the dataset for the post p. $T = P_s \cap R_s$. E_s is a set of extracted components from the dataset for the post p posted by the original author of the post. U_c is the total upvotes for comments proposing component c.

The **distinctiveness (Ds)** score is calculated as a percentage of unique components from all the components that the model generated:

$$Ds = \frac{|P_s|}{|P_{rs}|}$$

This score indicates the proportion of duplicates based on the matching algorithm. A higher distinctiveness score indicates greater originality in the generated components.

The **creativity (Cr)** score is calculated as a percentage of the components that are considered to be not included in manually extracted components from the Reddit comments. Formally, it can be calculated as

$$Cr = \frac{|P_s - R_s|}{|R_s|}$$

where and $|P_s - R_s|$ corresponds to the magnitude of set difference. This score evaluates the model's ability to generate components beyond the "core" set of responses present in the dataset. A higher Cr indicates an ability to create novel components.

The **upvote-weighted intersection (UWI)** score is calculated as a weighted average of upvotes for components from the set T. The score is calculated as follows:

$$UWI = \sum_{c \in T} \frac{U_c}{\sum_{k \in R_s} U_k}$$

This score reflects the importance of the predicted components in relation to how well they align with the opinions of Reddit users (indicated by the number of upvotes for the comments). A higher UWI value indicates better alignment between the model's predictions and the Reddit users.

The **crowd intersection (CI)** score is calculated as follows:

$$CI = \frac{|T|}{|R_s|}$$

This score is a percentage of components that appeared in both the model generated component set and the target dataset. Low CI indicates that the model generated a small amount of components that match the target components. Therefore, it missed a lot of components that were brainstormed in the discussion. High score indicates a high intersection of the model's outputs and the target human produced components.

The **final commit intersection (FCI)** score is calculated via the formula:

$$FCI = \frac{|P_s \cap E_s|}{|E_s|}$$

This score reflects the intersection (from CMRs perspective) of the components that the author of the post explicitly mentioned in their reply as doing or planning to do.

6 Experimental Setup

The experiments were conducted with the following LLMs: LLaMa3:8b, LLaMa3.1:8b (Dubey et al., 2024), and Gemma 2:9b (Team, 2024). The models were instructed with an initial system prompt explaining the task and outlining that the output format should clearly separate components. The model was prompted to generate a special stop token once it had finished generating the components. The experiments were conducted with $N = 0, 5$, and 10 examples from the dataset presented to the model. Each example included title, post content (without the author's update), and expected list of components with the stop token at the end. Then, the test title and post content were presented to the model. The model generated the components, and if the stop token appeared in the generated text, this was considered the final output. Otherwise, the model was presented with an additional request to generate more of components and complete the previous conversation history. This process was run until the stop token appeared in the text, or when the maximum allowed number of follow-ups (20) was reached.

For each of the few-shot experiments with $N \in \{5, 10\}$ examples the model was run independently 3 times, selecting random N examples from the dataset. The final metrics were averaged over the experiments per N. As a preprocessing step, all the generated components were run through the matching algorithm to remove duplicate components. To evaluate a joint performance of the models, the generated results per N were aggregated. We set random seed equals 2, and set other generation parameters to defaults, including a temperature of 0.7.

The constructed prompts and code are available in the project GitHub repository. The overview of the pipeline is presented in Figure 1.

7 Results

The results are presented in Table 2. The correlation plots of the metrics are presented in Appendix D.

A random sample of the 327 generated components was manually evaluated on their usefulness and relevance to the problem. The annotation results are presented in Figure 3 and in Appendix C. The models mostly generated useful components for tackling the input problem, with annotators' agreement on Useful/Not Useful labels reaching a Cohen Kappa score of 0.53. More details are provided in the Appendix B.

The correlation matrices of the scores are presented in Appendix D. Our results do not indicate a strong correlation between most metrics, except for the UWI and CI scores. This is expected, as CI is based on the intersection of LLM-generated and target components, while UWI reflects the importance of posted suggestions to Reddit users.

Based on the obtained results, all the models were able to output distinct sets of components when presented with examples. The distinctiveness scores in each run was 1.0 for N = 5 and 10. However, when models were not presented with examples from the dataset, LLaMa3.1 and Gemma 2 obtained Ds of 0.943 and 0.981 respectively. These scores are still high, but not as good as when presented with few-shot examples. In our experiments, the Ds score was not highly discriminatory, as all evaluated models demonstrated similar levels of competence.

The most creative model was LLaMa3:8b, as it was able to outperform other considered models with zero-shot (with Cr=1.557 and std of 1.048) and with N=10 (with Cr=1.574 and std of 0.275). When provided with 5 examples, LLaMa3.1 had the highest creativity score of 1.364. Not only was this result the best on average in the group of N=5, but it also was the most consistent one with std of 0.041.

On the other hand, when it comes to upvote-weighted intersection scores, there does not appear to be a clear winner. The UWI score can be interpreted as an approximation of utility of the predicted components based on the Reddit users' judgement. LLaMa3.1 achieved the best result with the zero-shot approach, with a score of 0.044 and the highest std of 0.14. In the 5-shot example exper-

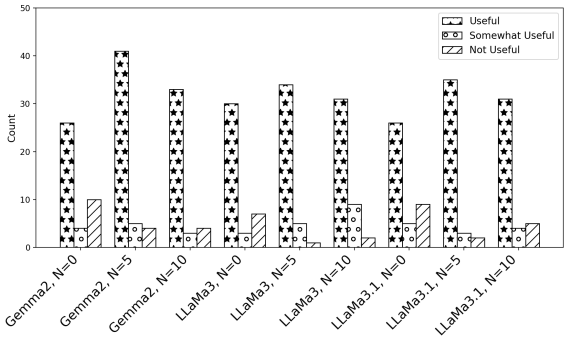

Figure 3: Distribution of useful, somewhat useful, and not useful components per model and per experiment.

Model	N	Ds *(std)*	Cr *(std)*	UWI *(std)*	CI *(std)*	FCI *(std)*
LLaMa3:8b	0	**1.0** *(0.0)*	**1.557** *(1.048)*	0.033 *(0.075)*	0.032 *(0.057)*	0.027 *(0.126)*
LLaMa3.1:8b	0	0.943 *(0.232)*	1.429 *(2.513)*	**0.044** *(0.14)*	**0.038** *(0.09)*	**0.061** *(0.194)*
Gemma 2:9b	0	0.981 *(0.136)*	1.299 *(1.029)*	0.023 *(0.078)*	0.016 *(0.04)*	0.021 *(0.121)*
LLaMa3:8b	5	**1.0** *(0.0)*	1.131 *(0.102)*	0.044 *(0.009)*	0.035 *(0.002)*	0.051 *(0.009)*
LLaMa3.1:8b	5	**1.0** *(0.0)*	**1.364** *(0.041)*	0.046 *(0.012)*	**0.041** *(0.002)*	**0.065** *(0.011)*
Gemma 2:9b	5	**1.0** *(0.0)*	1.134 *(0.113)*	**0.049** *(0.006)*	0.038 *(0.005)*	0.048 *(0.005)*
LLaMa3:8b	10	**1.0** *(0.0)*	**1.574** *(0.275)*	**0.055** *(0.002)*	**0.043** *(0.004)*	0.053 *(0.008)*
LLaMa3.1:8b	10	**1.0** *(0.0)*	1.295 *(0.039)*	0.053 *(0.006)*	0.037 *(0.002)*	**0.06** *(0.021)*
Gemma 2:9b	10	**1.0** *(0.0)*	1.123 *(0.043)*	0.042 *(0.01)*	0.033 *(0.004)*	0.039 *(0.013)*

Table 2: Average results per experiment for different LLMs on component generation task. N refers to a number of examples that was shown to the model. Ds stands for distinctiveness score, Cr is creativity score, UWI is upvote-weighted intersection score, CI is crowd intersection score, FCI is final commit intersection score. For N=0 only one experiment was conducted. In the brackets, *the standard deviation* is presented among the different runs. FCI was calculated only in the samples, where author provided indication of commitment to do a particular action.

iments, Gemma 2 performed better with an UWI of 0.049 and the lowest std of 0.006. Finally, in the 10-shot group, LLaMa3 showed a score of 0.055 with the lowest std of 0.002.

For crowd intersection score (CI), LLaMa3.1 outperformed other models in zero-shot and 5-shot settings with the scores of 0.038 and 0.041 respectively. In the 10-shot settings, LLaMa3 obtained the highest score of 0.043. These scores are quite low, indicating a small intersection with the components generated by the Reddit users' "brainstorm". Therefore, LLMs seemed to have missed a substantial proportion of possible components.

For the final commit intersection score (FCI), LLaMa3.1 outperformed other models in all experiments with the scores of 0.061 (zero-shot), 0.065 (5-shot), and 0.06 (10-shot). This score indicates an intersection with the "best" components - the ones that were selected by the original post author. However, in a lot of cases, this could also primarily represent a personal preference. Often, more context is required to determine what might be considered the best option for any particular problem.

In our experiments, we expected LLaMa3.1 to outperform LLaMa3 across the different experiments. However, LLaMa3 demonstrated the better performance in the N=10 settings. Similar behavior has been observed before. Based on the released results for these models, there are instances when LLaMa3:8b showed better results than LLaMa3.1:8b (for example, on GPQA (Rein et al., 2024) dataset, LLaMa3.1 obtained a score of 32.8 and LLaMa3 obtained 34.2 (Dubey et al., 2024)).

8 Conclusion

Our experiments showed that LLMs are capable of outputting distinct components for decision making. However, they still appear to be a far way from matching human judgement, even when presented with different examples of the expected alternative components. In our experiments, LLMs performed better when provided with more examples, as might be expected. In almost all the settings and experiments, the best performing models were LLaMa3 and LLaMa3.1. These models demonstrated the highest creativity, intersection with human judgement, and with which actions authors finally did or committed to doing. Nevertheless, intersection scores were overall still low, indicating room for improvement.

From a practical perspective, the creativity aspect is important in decision making as it provides a bigger picture for decision making. Generating alternatives automatically might allow decision makers to go beyond cognitive limitations. We showed that the considered LLMs are able to produce high creativity score, outputting possible components that were not considered in Reddit comments. Therefore, LLMs could be helpful in creating and extending lists of options which might serve as a starting points for decision makers to consider a problem from different perspectives.

Limitations

It is challenging to evaluate the generated components and their match to the actual target components. While we chose to utilize LLMs and man-

ually evaluated a sample from our experiments, further investigations might be required in order to create a more reliable metric. Similarly, newly generated solutions cannot be fully reliably evaluated from the utility perspective, though we gauged the usefulness of responses by manually evaluating a sample of generated components. We employed two annotators for the usefulness evaluation. However, the concept of usefulness is subjective and having only two evaluators may introduce bias and reduce statistical reliability. In future work, we aim to expand our pool of annotators and explore automated methods for evaluating usefulness.

In this work, we did not evaluate hallucination aspects of the models: LLMs are known to sometimes generate output unrelated to the topic. This is an important task the field might seek to address. Moreover, LLMs' inferences are consuming a lot of resources and time. Finally, the dataset we have can be extended further with more samples that include more diverse domains. However, considering the importance of competing alternatives in the decision making process, we believe that automatic alternative generation is a significant first step towards potential future computer-assisted decision making tools. Our experiments were conducted using smaller open-weight models with 7B and 9B parameters. In future work, we plan to explore larger models, including closed-source alternatives.

References

2017. First Quora Dataset Release: Question Pairs — quoradata.quora.com. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs. Online.

Hosam Al-Samarraie and Shuhaila Hurmuzan. 2018. A review of brainstorming techniques in higher education. *Thinking Skills and creativity*, 27:78–91.

Ernest R Alexander. 1979. The design of alternatives in organizational contexts: A pilot study. *Administrative Science Quarterly*, pages 382–404.

Ami Arbel and Richard M. Tong. 1982. On the generation of alternatives in decision analysis problems. *The Journal of the Operational Research Society*, 33(4):377–387.

John Beachboard and Kregg Aytes. 2013. An introduction to business problem-solving and decision-making. In *Proceedings of the Informing Science and Information Technology Education Conference*, pages 15–27. Informing Science Institute.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

E Downey Brill Jr, Shoou-Yuh Chang, and Lewis D Hopkins. 1982. Modeling to generate alternatives: The hsj approach and an illustration using a problem in land use planning. *Management Science*, 28(3):221–235.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Shoou-Yuh Chang, E Downey Brill Jr, and Lewis D Hopkins. 1983. Modeling to generate alternatives: A fuzzy approach. *Fuzzy Sets and Systems*, 9(1-3):137–151.

Alberto Colorni and Alexis Tsoukiàs. 2020. Designing alternatives in decision problems. *Journal of Multi-Criteria Decision Analysis*, 27(3-4):150–158.

Joseph F DeCarolis. 2011. Using modeling to generate alternatives (mga) to expand our thinking on energy futures. *Energy Economics*, 33(2):145–152.

Joseph F DeCarolis, Samaneh Babaee, Binghui Li, and S Kanungo. 2016. Modelling to generate alternatives with an energy system optimization model. *Environmental Modelling & Software*, 79:300–310.

Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Stanley D Fisher, Charles F Gettys, Carol Manning, Tom Mehle, and Suzanne Baca. 1983. Consistency checking in hypothesis generation. *Organizational Behavior and Human Performance*, 31(2):233–254.

Charles F Gettys, Rebecca M Pliske, Carol Manning, and Jeff T Casey. 1987. An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes*, 39(1):23–51.

Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual review of psychology*, 62(1):451–482.

Richards J Heuer. 1999. *Psychology of intelligence analysis*. Center for the Study of Intelligence.

Michael J Hicks. 1991. Brainstorming. In *Problem Solving in Business and Management: Hard, soft and creative approaches*, pages 87–107. Springer.

Raimo P. Hämäläinen, Tuomas J. Lahtinen, and Kai Virtanen. 2024. Generating policy alternatives for decision making: A process model, behavioural issues, and an experiment. *EURO Journal on Decision Processes*, 12:100050.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Thomas John. 2016. Supporting business model idea generation through machine-generated ideas: A design theory. In *ICIS*.

George Johnson et al. 1991. In the palaces of memory: How we build the worlds inside our heads. *(No Title)*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Helmut Jungermann, Ingrid Von Ulardt, and Lutz Hausmann. 1983. The role of the goal for generating actions. In *Advances in psychology*, volume 14, pages 223–236. Elsevier.

Ralph L. Keeney. 1992. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press.

Ralph L Keeney et al. 1994. Creativity in decision making with value-focused thinking. *Sloan Management Review*, 35:33–33.

L Robin Keller and Joanna L Ho. 1988. Decision problem structuring: Generating options. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(5):715–728.

Jacob Kohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Antonio Leal, Steven Levin, Steven Johnson, Marcy Agmon, and Gershon Weltman. 1978. An interactive computer aiding system for group decision making. Technical report, Tech. Rep. PQTR-1046-78-2.

Antonio Leal and Judea Pearl. 1977. An interactive program for conversational elicitation of decision structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(5):368–376.

Orfelio G León. 1999. Value-focused thinking versus alternative-focused thinking: Effects on generation of objectives. *Organizational Behavior and Human Decision Processes*, 80(3):213–227.

Holy Lovenia, Rahmad Mahendra, and et al. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *arXiv preprint arXiv: 2406.10118*.

Matteo Morelli, Maria Casagrande, and Giuseppe Forte. 2022. Decision making: A theoretical review. *Integrative Psychological and Behavioral Science*, 56(3):609–629.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Vladimir M Ozernoy. 1985. Generating alternatives in multiple criteria decision making problems: A survey. In *Decision Making with Multiple Objectives: Proceedings of the Sixth International Conference on Multiple-Criteria Decision Making, Held at the Case Western Reserve University, Cleveland, Ohio, USA, June 4–8, 1984*, pages 322–330. Springer.

Judea Pearl, Antonio Leal, and Joseph Saleh. 1982. Goddess: A goal-directed decision structuring system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3):250–262.

Gordon F Pitz, Natalie J Sachs, and Joel Heerboth. 1980. Procedures for eliciting choices in the analysis of individual decisions. *Organizational Behavior and Human Performance*, 26(3):396–408.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. Ai-augmented brainwriting: Investigating the use of llms in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Herbert A. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Randall Steeb and Steven C Johnston. 1981. A computer-based interactive system for group decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(8):544–552.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Dataset Creation Process

In this section, we illustrate the dataset creation process. Each annotator is presented with the title of a Reddit post along with two related comments (see Figure 4).

The annotation process begins with the annotator reviewing the post to determine its suitability for the task. If the post is deemed relevant, the annotator initializes an empty List of Extracted Components for the post (LEC).

Next, the annotator examines each comment. If a comment contains a relevant component, the annotator checks whether this component has already been extracted for the post by reviewing the LEC. If the component is already in the LEC, the annotator assigns the existing component to the comment's ID. Otherwise, if the component is not in the LEC, the annotator extracts a clear version of the component from the comment, adds it to the LEC, and assigns it to the comment's ID.

For example, consider the post shown in Figure 4. When annotating the comment with ID coc83mo, the annotator extracts the component: "Take your children to have their vaccines updated behind your husband's back.".

This component is added to the LEC and mapped to coc83mo. Next, the annotator reviews the comment with ID coc1bd9. After comparing it against the LEC, they determine that no existing components match according to the Component Matching Rules (CMR). Therefore, a new component is extracted: "Talk to your husband and try to come up with a compromise—such as vaccines that do not contain a specific ingredient.". This new component is added to the LEC and assigned to coc1bd9. This process is repeated for all comments associated with the post. Once all comments are annotated, the LEC is reset, and a new LEC is initialized for the next post.

B Components Matching Algorithm Evaluation and Architecture

B.1 Algorithm Architecture and Manual Evaluation

To determine whether the pair of components are identical with respect to the defined CMRs, we developed an LLM-based ensemble method, utilizing LLaMa3:8b (Dubey et al., 2024) and Mistral:7b (Jiang et al., 2023) models[4].

The components matching algorithms was designed as follows. Firstly, LLaMa3 is provided with a prompt which provides the model with a set of CMRs. A few examples were provided as well, covering all of the rules. These examples were created manually outside of the dataset with the total number of examples of 4. All prompts and instructions are available in the project GitHub repository. The model then is instructed to predict "MATCH" (if the pair of components are the same) or "NOT MATCH" (if at least one of the CMRs does not hold). Where LLaMa3 fails to output the suitable value, the same inputs are provided to the Mistral model. If Mistral is not able to output the result, "NOT MATCH" label is assigned. During testing, models were able to output a value in the expected format for all the samples. We did not consider embeddings similarity-based techniques (e.g. cosine similarity-based using transformers or sentence similarity pre-trained models) as they are not able to match a specific set of rules, but only consider the overall semantics of the sentence

[4]We experimented with other LLMs, but this combination provided the best overall result. Embeddings similarity-based techniques (e.g. cosine similarity-based using transformers or sentence similarity pre-trained models) were not able to match a specific set of rules, but only considered the overall semantics of the inputs.

SUBMISSION:
2uvcdq

Should I to take my children to have their vaccines updated behind my husband's back?
****
I know Reddit generally comes off as pro-vaccine and will hopefully have my back on this. There is no way to convince him to agree with me or to change his mind. A little back story as to why they're not up to date: My kids are 6 and 7. My youngest began talking rapidly and early. When she was two, we had a dose of MMR--she stopped talking dead in her tracks. At that point, both kids were up to date. She had speech therapy for almost a year and is fine now, but it was a struggle. This was the defining moment for my husband to decide no more vaccines for either. I'm not going to lie, as a Mother, it definitely freaked me out--but I feel that they're in the safe zone and need these shots before they get kicked out of school...or sick...or die. The only shot they did not get, aside from boosters, is the Polio vaccine. My husband has agreed to this. I want them to have their missing boosters, however, and as much as I hate to go behind his back and do it, they're my children too! I just really need some support to justify my actions because, as I said, he's not going to see eye-to-eye with me ever. I feel like this is my only option at this point. I know some people might think it's wrong that I would go behind his back, but isn't it just as wrong that he refuses to do what I feel is right as well? Thanks Reddit!
==========
COMMENT
coc83mo
I fully support doing it behind his back.
---
coc1bd9
Maybe a compromise? The offending ingredient that sets people off (for whatever reason) seems to be Thimerosol. Maybe find a place that has the shots without that ingredient?

Figure 4: Example of the sample in the annotation system.

	Precision	Recall	F1
MATCH	0.956	0.886	0.919
NOT MATCH	0.987	0.995	0.991
Macro avg	0.971	0.940	0.955
Weighted avg	0.984	0.984	0.984

Table 3: Results of the ensemble matching metric.

input.

To evaluate the proposed algorithm, we made use of the new dataset by gathering all unique individual components from the labeled Reddit dataset. As per the annotation guidelines, all the components per post are considered to be unique, i.e., the same advice suggestions were always summarized in the same way. Hence, we could derive a set of gold-standard "NOT MATCH" samples from all combinations of any two unique components per post. The total number of negative ("NOT MATCH") samples was 10,841. The positive ("MATCH") examples were derived by paraphrasing all unique components from the dataset with Mistral using a zero-shot approach. The model was provided with the instruction to preserve required components, as was described in the previous paragraph. The paraphrased versions were manually reviewed afterwards to ensure that the paraphrase fit the requirements. As a result, 1,184 samples were accepted and 181 rejected.

Finally, the proposed approach was run on the combined dataset of "MATCH" and "NOT MATCH" pairs. The results are presented in Table 3.

Considering that all metrics exceed 90%, particularly the recall metric for "NOT MATCH" class, the ensemble approach is effective for determining whether components are the same. While the metrics are not perfect, they apppear reasonable and demonstrate strong performance.

B.2 Manual Evaluation of Generated and Manually Extracted Components

After the LLMs generated components, we took a random 300 of pairs of generated and manually extracted components (71 was marked as matching and 229 as not matching by the algorithm). The pairs were selected randomly across all experiments (see distribution on the Table 4). Annotator a manually evaluated those pairs. The results are presented in Table 5. The overall accuracy is 0.92. Results indicate a high agreement, therefore the algorithm could be considered reliable.

	N. Examples	**Run**	**N. Samples**
Gemma2	0	1	10
Gemma2	10	1	12
Gemma2	10	2	11
Gemma2	10	3	23
Gemma2	5	1	21
Gemma2	5	2	15
Gemma2	5	3	12
LLaMa3	0	1	14
LLaMa3	10	1	20
LLaMa3	10	2	15
LLaMa3	10	3	17
LLaMa3	5	1	13
LLaMa3	5	2	12
LLaMa3	5	3	14
LLaMa3.1	0	1	9
LLaMa3.1	10	1	17
LLaMa3.1	10	2	7
LLaMa3.1	10	3	16
LLaMa3.1	5	1	9
LLaMa3.1	5	2	13
LLaMa3.1	5	3	20

Table 4: Number of evaluated pairs per model and per run.

	Prec	**Rec**	**F1**	**Num**
MATCH	0.77	0.96	0.86	71
NOT MATCH	0.99	0.91	0.95	229
Macro Avg	0.88	0.94	0.90	300
Weighted Avg	0.94	0.92	0.93	300

Table 5: Results of manual evaluation of generated and manually extracted components matching. *Prec* refers to the precision score. *Rec* refers to the recall score.

C Usefulness Evaluation of Generated Components

Additionally, we manually evaluated the usefulness/relevance of the model generated components. The same annotators a and b were recruited to annotate the sample from the pool of generated components. The components were selected evenly across the models, number of few-shot examples shown to the model, and experiment runs. The annotators were shown an original Reddit post and a generated component. They had to determine, if the component is relevant to the post and if it is useful (assign a label **U**), somehow useful (assign

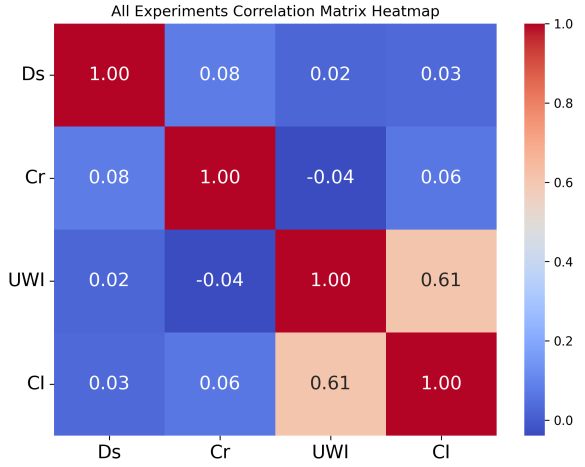

Figure 5: Correlation plot of aggregated results of the experiments over all runs.

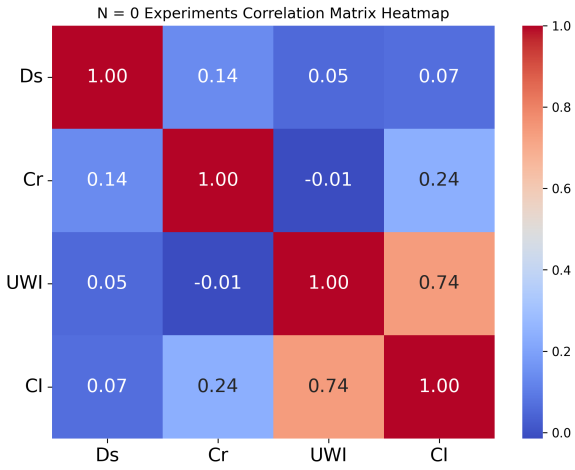

Figure 6: Correlation plot of aggregated results of the experiments with N=0 over all runs.

a label **SU**) or not useful and/or irrelevant (assign a label **NU**). The annotation results are presented in Table 6.

To address generalization of the experiment results, we employed the following strategy to calculate the annotators' agreement. Given the subjective nature of the task, we merged "Useful" and "Somehow Useful" labels into a single category and calculated the Cohen's Kappa score based on the intersection of 31 samples. The obtained score was 0.53, indicating moderate agreement.

D Metrics Correlation

The correlation plots are presented on Figures 5,6,7, and 8. Out results show that in majority of cases models predicted relevant and useful components to the presented problem. As in some of our experiments, distinctiveness score (Ds) did not have variation, its values are missing. Our findings show that the upvote-weighted intersection score (UWI) has correlates with the crowd intersection score (CI). It is expected due to a design of these metrics: they both are based on the intersection of the generated components and manually annotated components that are matched to them. Other pairs of metrics do not show high correlations.

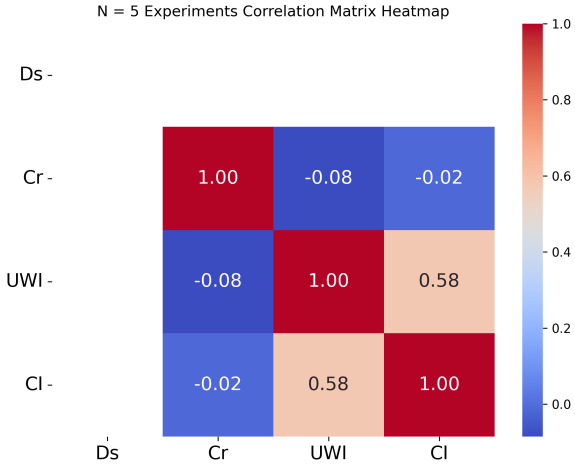

Figure 7: Correlation plot of aggregated results of the experiments with N=5 over all runs.

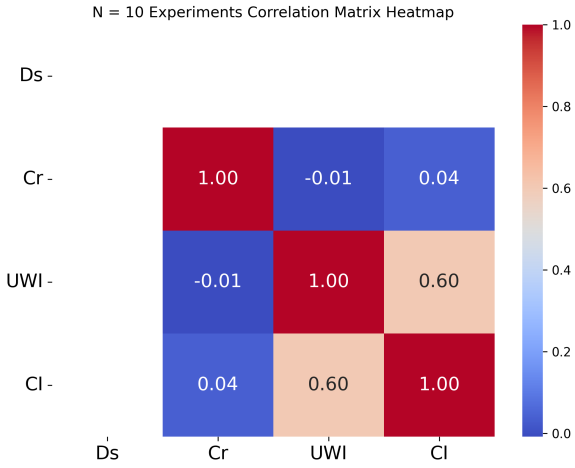

Figure 8: Correlation plot of aggregated results of the experiments with N=10 over all runs.

	Num. U			Num SU			Num NU		
	Ant. a	*Ant. b*	*Total*	*Ant. a*	*Ant. b*	*Total*	*Ant. a*	*Ant. b*	*Total*
Gemma2, N=0	24	2	**26**	3	1	**4**	10	0	**10**
Gemma2, N=5	30	11	**41**	2	3	**5**	2	2	**4**
Gemma2, N=10	21	12	**33**	1	2	**3**	2	2	**4**
LLaMa3, N=0	28	2	**30**	1	2	**3**	5	2	**7**
LLaMa3, N=5	22	12	**34**	3	2	**5**	1	0	**1**
LLaMa3, N=10	20	11	**31**	1	8	**9**	0	2	**2**
LLaMa3.1, N=0	23	3	**26**	3	2	**5**	7	2	**9**
LLaMa3.1, N=5	24	11	**35**	1	2	**3**	1	1	**2**
LLaMa3.1, N=10	20	11	**31**	0	4	**4**	2	3	**5**

Table 6: Results of manual evaluation of usefulness and relevance of the generated components. **U** indicates *Useful and Relevant*, **SU** indicates *Somehow Useful and Relevant*, and **NU** indicates *Not Useful and/or Irrelevant*. **N** indicates a number of few-shot examples. *Ant. a* refers to the results by the annotator a, and *Ant. b* refers to the results by the annotator b.