Framework for Computational Component-Based Alternatives Generation in Decision-Making

Anonymous ACL submission

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 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' comments. We also introduce new metrics to assess the quality of generated alternatives, including distinctiveness, creativity, upvote-weighted, crowd-intersection, and did-intersection scores. As a baseline, we evaluated the large language models (LLMs) LLaMa3:8b, LLaMa3.1:8b, and Gemma 2:9b on this 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. 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 actions.

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. This paper introduces component-based alternatives generation with the aim of providing first steps towards aiding the process of human alternative generation using machine learning.

Different theories disagree on the exact structure of the decision-making process (Morelli et al., 2022). However, most theoretical frameworks consider decision-making to be the process of selecting an action from a set of alternative options, frequently without specifying from where this set of alternatives originates. Some studies (Hämäläinen et al., 2024; Pitz et al., 1980; Jungermann et al., 1983; Starr and Greenwood, 1977) discuss potential cognitive processes involved in creating a list of alternative options, but there still is an absence of computational resources for modeling decision-making through alternative generation.

It has been highlighted in multiple studies that the process of creating a starting point in form of a set of possible actions is a very important yet challenging step (Hämäläinen et al., 2024; Fisher et al., 1983). To create a list of possible actions, people need to access all relevant information in their long-term memory, although this information may not be organized in a coherent structure useful for the problem (Jungermann et al., 1983). Other information can be added to the existing structure by combining various existing knowledge and new information (Einhorn and Hogarth, 1981). Finally, knowledge can be represented using an associative network (Huggett et al., 2007), where nodes are cognitive units and arcs that represent connections between them (Anderson, 2013), creating a complex hierarchical system. When searching for an answer, cognitive units (e.g. prototypes of sets of objects, causal patterns, and spatial images) are selectively activated based on their relevance to the current problem (Johnson-Laird, 1980). Activation spreads from node to node, following nodes with more direct links until the final "goal" node is extracted (McClelland and Rumelhart, 2020).

As such, memory retrieval of relevant information is a key process in decision-making (Johnson et al., 1991). The more frequently a pattern is observed or referred to, the stronger is the connec-

tion and the easier it is to extract information from there. At the same time, interference from previous knowledge can also make it more difficult to restructure information to see the problem from different perspective (Heuer, 1999).

In familiar, well-defined problems, option generation is often routine, as people rely on problem prototypes and other heuristics recalling familiar options (Gigerenzer and Gaissmaier, 2011). This process helps us generate options based on similar past experiences, for example by learning from previous mistakes. However, it can also lead to biases such as an increased difficulty to form new connections, a tendency to learn towards highly representative options, overlooking less familiar alternatives, and struggling to retrieve or creatively generate high-quality solutions (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 judgement 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. Overcoming humans' limited cognitive flexibility, "brainstorming" is also particularly well-suited for computational automation (John, 2016).

Developing an automatic algorithm could help overcome above-mentioned biases by generating wider set of possible alternatives. Such algorithms can 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: given a decision-making problem, the algorithm needs to generate a list of possible actions to solve it. As a baseline, we evaluate an algorithm designed to automate the generation of alternatives for decision-making problems by utilizing large language models (LLMs). To evaluate the performance of the models, a new dataset based on Reddit comments was manually labeled, approximating the "brainstorm" technique by incorporating advice from multiple users per question. The following contributions are made:

- We propose a new definition of alternative actions (alternative options) based on atomic actions (components), and introduce tasks for generating alternatives. Atomic actions are the minimal actions that can be taken to solve the task. Alternative options could include a set of possible atomic actions.

- We present a new dataset for the proposed tasks based on the Reddit Advice subreddit[1] (r/Advice). Specifically, we filtered posts including requests for 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: **creativity**, **upvote-weighted intersection**, **crowd-intersection**, and **did-intersection** scores. These metrics leverage an ensemble of large language models-based (LLMs) matching algorithm that checks if atomic actions imply the same. The matching algorithm was manually validated.

- We evaluate different LLMs for alternative generation using both zero-shot and few-shot approaches. For few-shot, 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 et al., 2024).

2 Related Work

No dataset or evaluation metric 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 and in particular domains. A comprehensive survey and overview of different

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

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), the author introduced a "design" concept prior the the "choice".

MGA (Modelling 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 optimisation algorithms to explore the neighbourhood 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 underresearched 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; 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). Lastly, 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.

3 Definitions and Tasks

In this work, an *alternative option* (AO) refers to an action or a set of actions that a person could take in the context of a given problem. Each AO consists of smaller units of action, referred to as *components*, which are atomic, i.e., should not be broken down further. More formally, components are the smallest actions that can be taken to achieve a result in solving the problem. They may include conditional elements, a certain order of actions, concretize specific actors who should perform the action, 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 include actions such as:

1. Do not tell B that you took the cookie.
2. Tell B that you took the cookie.
3. Buy B a whole pack of cookies as a means of apology.
4. If you value B's friendship, tell the truth.
5. Take more cookies.
6. Tell B that you took the cookie and, if B is angry, buy B a lunch.
7. Buy B a lunch.

In the example, there are multiple suggested component actions that A can take. Some of them can be taken together (e.g., number 1 and 5). If the components are mutually exclusive (e.g. number 1 and 2), they are referred to as *competing components*. An 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.

Following this logic, the task of alternative generation can be structured as follows: (i) **Component Generation (CG)** and **Component Competition (CC)**.

The 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.

The CC task is a binary classification problem: given two components, the algorithm must determine whether they are competing.

In this paper, we focus on the CG task.

4 Dataset

4.1 Annotation and Agreement

In this work, we introduce a novel dataset which was specifically created to fit the task. For this, we manually labeled the Reddit Corpus, a dataset provided in 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 University of Dundee (UK).

The annotators were presented with a 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: 50 most upvoted comments, 25 random comments and pairs of comments that the author of the post responded to (comment and author reply). The annotators were asked to read the post and determine if in the post the problem was stated clearly and the author was asking for help in a search of possible actions or options to take. All of the 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 also disregarded. If the post was considered to fit 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. As such, alternatives can be said to constitute sets of components identified in each comment. 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[2].

After the annotation was complete, the selected posts were reviewed again and "update" parts of the post (additional text that had been added after the discussion had formally ended) was removed. As it usually had not been available to commentators when they proposed their solutions, the update part was counted as one of the author responses and labeled as such.

To measure inter-annotator agreement, the following strategy was deployed. The 15 same posts were included in the labeling chunks for both of the annotators and included 148 comments in total. The total numbers of comments that were considered to have at least one component were 79 and 120 for annotators a and b respectively. The total numbers of unique components for all the posts were 49 and 51 for annotators a and b respectively. The average alternative numbers per post were -0.803 and -0.792 for annotators a and b respectively. The Cohen's kappa score (Kohen, 1960) was calculated for integer value of number of components that were extracted from the comment and its value is 0.614, indicating substantial agreement.

4.2 Dataset Statistics

After filtering the posts that did 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

[2] HIDDEN FOR REVIEW

	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.

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 author appeared to commit to take some of the actions is 75 (70.7% of all unique posts), with total number of unique components of 197 and number of comments of 240 (sometimes, author replied to multiple comments with the same commitment).

The total number of words can be found in the 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 **did-intersection** scores.

All the proposed metrics require an algorithm that determines whether the two components are identical or not (in case of intersection, duplicate removal etc.). Components are considered identical if they fit the following **component matching rules (CMRs)**: if they preserve the order of actions and overall semantics, maintain the same conditional parts, and refer to the same entities or people. 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 LLaMa3:8b (Dubey et al., 2024) and Mistral:7b (Jiang et al., 2023) models[3].

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.

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 2.

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

[3]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.

	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 2: Results of the ensemble matching metric.

demonstrate strong performance.

5.2 Metrics

Let

- P_{rs} be a set of all predicted components (possible repetitive) and P_s a set of predicted *unique, none-repetitive* components (based on the matching algorithm and CMRs) from the model for the post p.
- $R_s = \{c_i | i = 1, \ldots, N\}$ be a set of extracted components from the dataset for the post p.
- $T = P_s \cap R_s$.
- D_s be a set of extracted components from the dataset for the post p, which the author of the post performed.
- U_c be the total upvotes for comments proposing component c.

The **distinctiveness (Ds)** score is calculated as a percentage of unique components to all the components that 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 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 components from the set T based on the upvotes they received. For each component, the averaged upvote score is assigned as

$$NU_c = \frac{U_c}{sum_{c \in R_s} U_c}$$

The score is calculated as follows:

$$UWI = sum_{c \in T} NU_c$$

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.

The **did-intersection (DI)** score is calculated via the formula:

$$DI = \frac{|P_s \cap D_s|}{|D_s|}$$

This score reflects a percentage of components that appeared in the generated component set and components that author ended up doing or commited to doing.

6 Experimental Setup

The experiments were conducted with the following LLMs: LLaMa3:8b (Dubey et al., 2024), LLaMa3.1:8b (Dubey et al., 2024), and Gemma 2:9b (Team et al., 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 asked 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, it was considered as 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 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 the duplicate components. To evaluate a joint performance of the models, the generated results per N were aggregated.

The constructed prompts and code are available in the project GitHub repository.

7 Results

The results are presented in the Table 3.

Based on the obtained results, all the model 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 the examples from the dataset, LLaMa3.1 and Gemma 2 struggled with outputting distinct sets with scores of 0.943 and 0.981 respectively. These scores are still high, but not as good as when presented with few-shot examples.

The most creative model was LLaMa3:8b, as it was able to outperform other considered models with the zero-shot approach (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 experiments, 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 respectivelly. 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 did-intersection score (DI), 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 had been 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.

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..

Model	N	Ds *(std)*	Cr *(std)*	UWI *(std)*	CI *(std)*	DI *(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 3: 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, DI is did-intersection score. For N=0 only one experiment was conducted. In the brackets, *the standard deviation* is presented among the different runs. DI was calculated only in the samples, where author provided indication of commitment to do a particular action.

9 Limitations

It is challenging to evaluate the generated components and their match to the actual target components. While we chose to utilize LLMs, further investigations might be required in order to create a more reliable metric. In this work, we did not consider potential utility or 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' inference is consuming a lot of computational resources and time. The dataset we have can be extended further with more samples that include more diverse domains. Finally, we chose not to evaluate the quality of the generated components. This task is difficult to formalize, and reasonable datasets and metrics are likewise challenging to develop. However, considering the importance of competing alternatives in the decision-making process, we believe that automatic alternative generation is a significant first steps towards potential future computer-assisted decision-making tools.

Acknowledgments

References

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.

John R Anderson. 2013. *The architecture of cognition*. Psychology Press.

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.

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, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,

Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev,

Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hillel J Einhorn and Robin M Hogarth. 1981. Behavioral decision theory: Processes of judgment and choice. *Journal of Accounting Research*, pages 1–31.

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.

Michael Huggett, Holger Hoos, and Ron Rensink. 2007. Cognitive principles for information management: The principles of mnemonic associative knowledge (p-mak). *Minds and Machines*, 17:445–485.

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)*.

Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science*, 4(1):71–115.

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.

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.

James L McClelland and David E Rumelhart. 2020. Distributed memory and the representation of general and specific information. In *Connectionist Psychology*, pages 75–106. Psychology Press.

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.

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.

Martin K Starr and Leonard H Greenwood. 1977. Normative generation of alternatives with multiple criteria evaluation. *Multiple criteria decision making. New York: North-Holland*, pages 111–28.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.