The Effect of Question Formats in Dialogue with Generative AI: A Comparative Analysis of Open Questions and Prompt Engineering

Keisuke Sato[1*]

[1*]Natural Science, National Institute of Technology, Ibaraki College, Nakane, Hitachinaka, 312-8508, Ibaraki, Japan.

Corresponding author(s). E-mail(s): skeisuke@ibaraki-ct.ac.jp;

Abstract

As dialogue with generative AI enters a new phase, this study compares and analyzes the effects of question formats in AI interactions, specifically comparing open-ended questions with prompt engineering (structured questions). We conducted experiments using eight AI models in four areas: the future of education, improving communication within companies, family meal planning, and new movies, and analyzed them using 20 evaluation metrics. The results showed that the question format had a multifaceted and complex effect on the AI response. Open questions showed an advantage in terms of creativity, diversity, and thought promotion, while prompt engineering formats were effective in terms of concreteness and naturalness of dialogue. In addition, the effectiveness of the question format was highly dependent on the theme and AI model used. These findings suggest the importance of strategic selection of question formats in designing interactions with AI. Choosing the appropriate question format according to the purpose, theme, and AI model used can lead to more effective and creative interactions. This research provides new guidelines for effective interaction with AI and emphasizes the importance of improving AI literacy.

Keywords: Generative AI, Question formats, Open-ended questions, Prompt engineering, Human-AI interaction

1 Introduction

The rapid advancement of generative AI based on large-scale language models (LLMs) has ushered in a new era of human-AI interaction, offering unprecedented capabilities in complex problem-solving and creative tasks [1, 2]. These sophisticated computational intelligence systems have far surpassed traditional information retrieval methods. However, despite their potential, the optimal strategies for leveraging these AI systems remain largely unexplored [3], presenting a critical challenge in the field of computational intelligence.

Prompt engineering has emerged as a widely recognized method for interacting with AI, playing a crucial role in unlocking the potential of these systems [4]. This approach typically involves providing clear and specific instructions to AI models [5]. However, from the perspective of maximizing the capabilities of computational intelligence systems, prompt engineering is still in its infancy. Recent research has focused on developing more effective prompting techniques, such as the catalog of prompt patterns in software engineering proposed by White et al. [6, 7]. While these efforts have shown promise, they are often limited by their focus on specific domains and may not generalize well to broader applications of computational intelligence.

In contrast, open-ended questions (defined as questions that do not limit the response and are intended to elicit free thinking and diverse perspectives) are known to encourage free thinking in respondents and elicit diverse perspectives. De Wit (2020) showed that open-ended questions can be a useful means of gaining a more flexible understanding of students' views of the future[8]. Extending this concept to AI interaction, the ability to generate and evaluate novel ideas becomes crucial for AI systems to engage in meaningful and creative dialogues. The Ul Haq et al. (2024) paper explores the application of sentence embedding models for evaluating the novelty of ideas generated in open-ended, collaborative problem-solving scenarios, which aligns well with the concept of encouraging free thinking and diverse perspectives in AI interactions[9]. The Ecoffet et al. (2020) paper, on the other hand, delves into the safety aspects of open-ended AI, raising concerns about the potential unpredictability and risks associated with highly creative AI systems[10]. It emphasizes the need for careful consideration of the balance between fostering creativity and maintaining control, which is crucial when extending open-endedness to AI interactions to ensure safe and productive outcomes.

The dichotomy between structured prompt engineering and open-ended questioning represents a fundamental challenge in computational intelligence system design. While traditional prompt engineering methods have focused on controlling AI output through specific instructions, emerging approaches like the Tree of Thoughts (ToT) encourage more open-ended thinking from AI systems [11]. However, these newer methods, while innovative, have not been systematically compared with traditional approaches across a wide range of tasks and AI models.

Despite the growing importance of this topic, there has been no direct comparative study of open-ended question formats versus prompt-engineering formats in interactions with computational intelligence systems. This research gap is particularly significant given the potential of generative AI in collaborative and creative problem-solving [12], enhancing human creativity [13], improving business models [14], and advancing research methodologies [15]. The lack of comprehensive comparative studies hinders our ability to design optimal interaction strategies for these powerful computational intelligence tools.

The purpose of this study is to address this critical gap by conducting a comprehensive comparison and analysis of the effectiveness of open-ended questions versus prompt-engineering questions across multiple AI models and various domains. We hypothesize that:

1. Open-ended questioning techniques will enhance the creativity and diversity of AI responses compared to traditional prompt engineering methods.
2. The effectiveness of question formats will vary depending on the specific AI model and the domain of application.
3. A hybrid approach combining elements of both open-ended and prompt-engineering techniques may yield superior results in certain contexts.

Our research makes several novel contributions to the field of computational intelligence:

1. We provide a multifaceted and quantitative evaluation of the effects of different question formats on AI responses, analyzing eight state-of-the-art AI models across four diverse themes using 20 distinct evaluation metrics.
2. We offer new insights into the interaction between AI model characteristics and question formats, which is crucial for optimizing human-AI collaboration in computational intelligence systems.
3. We propose guidelines for strategically selecting question formats based on the specific goals, themes, and AI models used, potentially leading to more effective and creative interactions with computational intelligence systems.
4. We contribute to the improvement of AI literacy by elucidating the importance of questioning techniques in AI interactions, which is essential for the broader adoption and effective use of computational intelligence technologies.

This research is expected to have significant implications for the design of more effective human-AI interaction paradigms. By systematically comparing different questioning techniques across various AI models and domains, we aim to provide a foundation for enhancing creative problem-solving capabilities, streamlining knowledge creation processes, and improving educational and business applications of computational intelligence systems.

In the following sections, we will detail our experimental design, including the selection of AI models, the development of our evaluation metrics, and our data analysis methodologies. We will then present our findings, discuss their implications for the field of computational intelligence, and propose directions for future research in this rapidly evolving domain. Through this comprehensive analysis, we aim to bridge the gap between theoretical understanding and practical application of questioning techniques in computational intelligence systems, paving the way for more effective and innovative human-AI collaboration.

2 Experimental Design

The primary objective of this study is to compare and analyze the effects of question formats in dialogues with generative AI. Specifically, we address the following research questions:

1. Elucidate the differences in effectiveness between open-ended questions and prompt engineering questions.
2. Examine how the effects of question formats vary depending on AI models and themes.
3. Evaluate how question formats influence the quality and effectiveness of AI responses.

To address these questions, we adopted the following experimental design.

2.1 Themes and Question Types

We selected four themes: the future of education, improving internal corporate communication, healthy family meal planning, and new movies. These themes represent areas where AI dialogue is useful and diverse perspectives are required. For each theme, we prepared two types of questions: open-ended questions and prompt engineering questions. The actual questions used are provided in the appendix A.1.

2.1.1 Open-ended Questioning Technique

The "Interactive Ideation Approach" proposed in this study is a new methodology designed to facilitate creative dialogue with AI. This technique has the following characteristics:

1. Provision of rich context
2. Externalization of the thought process
3. Multi-layered information blending
4. Minimize constraints
5. Collaborative Exploration
6. Creative Thinking

2.1.2 Prompt engineering style questions

Prompt engineering style questions were created in accordance with the Claude prompting guide.md[16], a document provided by Claude, the AI assistant from Anthropic.

2.2 AI Models

The following AI models were used:

1. Coral (Command R+) (A)
2. ChatGPT 4.0 Turbo (B)
3. Gemini 1.0 Pro (C)

4. Gemini 1.5 Flash (D)
5. Gemini 1.5 Pro (E)
6. Claude-3-haiku-20240307(F)
7. Claude-3-opus-20240229(G)
8. Claude-3-sonnet-20240229(H)

These models were selected based on their recognition in Japan and Japanese processing ability.

In this study, all questions and answers were conducted in English for the following reasons:

- Ease of text mining
- Universality of language
- Fair comparison between models

50 responses were generated for each model and each question. This number was chosen as a sample size to obtain statistically significant results. Data collected from 3 July 2024 to 15 July 2024.

2.3 Analysis Method

2.3.1 Evaluation Metrics Using Open Source Tools

We used some open-source tools[17] to calculate metrics such as text length (Length), Gunning Fog Index, Rix, and Measure of Textual Lexical Diversity (MTLD). These metrics allow us to analyze from multiple perspectives how the question format affects the complexity of the model response, its readability, and the diversity of its vocabulary.

2.3.2 Original evaluation metrics

In order to gain a deeper understanding of the responses of AI models, we defined 16 original evaluation metrics in addition to the open-source metrics shown in 2.3.1. These metrics quantify the qualitative aspects of responses, such as creativity, practicality, concreteness, naturalness of dialogue, and promotion of thinking. Detailed definitions and calculation methods for each indicator are provided in the appendix A.1.

2.3.3 Data Preprocessing

We used the Python libraries NLTK, spaCy, and TextBlob for data preprocessing. The preprocessing steps included the following:

1. Tokenization
2. Stop word removal
3. Lemmatization

2.3.4 Extraction and Analysis of Frequent Words

1. The top 20 frequent words are extracted for each model and each theme.

2. From 8 models $\times$ top 20 words = 160 words, words that appear five or more times are identified.
3. From these words, we extract words that are biased towards either Answer 1 (open-ended question format) or Answer 2 (prompt engineering format).

2.3.5 Evaluation Metrics

In this study, we defined 16 evaluation metrics to evaluate the quality of ideas generated by AI models from multiple perspectives. These metrics aim to quantify various aspects of ideas and to quantitatively analyze the impact of question types.

The main evaluation metrics are as follows:

1. Creativity
2. Practicality
3. Specificity
4. Interactive nature
5. Thought-provoking

Other evaluation metrics include complexity, technicality, diversity, consistency, readability, density of proper nouns, density of parts of speech, average word length, lexical diversity, dependency distance, and frequency of passive voice usage.

These metrics are automatically calculated from the text data of ideas using natural language processing methods. Detailed definitions, calculation methods, and explanations of the implementation of each indicator are provided in the appendix.

2.3.6 Normalization and Comparison of Scores

Min-Max scaling was used to normalize the evaluation scores of each idea to a range of 0 to 1. This allows us to compare different evaluation metrics. The mean and standard deviation of the scores for each file were calculated, and bar charts were used to visualize the distribution of scores.

2.3.7 Statistical Analysis

In this study, we conducted a pairwise t-test to statistically analyze the differences in responses to open-question and prompt-engineering formats for each AI model for both the evaluation metrics using open-source tools and our own metrics. In this analysis method, we compared the file pairs of Answer 1 (open-ended question format) and Answer 2 (prompt engineering format) for each model, and conducted a test for all evaluation metrics.

The results of the t-test are reported in the form of t-statistics and p-values. If the p-value is less than 0.05, we determined that there is a statistically significant difference between the two question types for that evaluation metric. The sign of the t-statistic indicates whether the open-ended or prompt-engineering question type is dominant.

This analysis method allowed us to obtain more detailed and reliable findings about the characteristics of each AI model and how the question type affects the quality of the answers. It also allowed us to gain deeper insights into the comparison between models and the relationship between evaluation metrics.

2.4 Limitations and Potential Biases

This study may have the following limitations and potential biases:

- Language bias due to the use of English only
- Bias related to selected themes and AI models
- Limitations on generalizability due to sample size and experimental period constraints
- Potential bias in the education theme questions due to the author's background as a teacher

We need to be cautious in interpreting and generalizing the results, recognizing these limitations.

2.5 Positioning in Computational Intelligence Systems

This study makes an important contribution to the field of computational intelligence systems by exploring effective methods of dialogue with generative AI. In particular, by quantitatively analyzing the impact of question formats on AI responses, it provides insights for designing more effective human-AI interactions. This plays an important role in expanding the applicability of AI in various fields such as education, business, and creative work.

3 Experimental Results

3.1 Frequent Word Analysis

A list of frequently occurring words (top 20 words) extracted from the responses of the eight AI models for each theme is given in the Appendix15, 17, 19, 21. For each theme, frequent words were extracted and compared from responses to open-ended questions (Answer 1) and prompt engineering questions (Answer 2)1. Additionally, for the highest-scoring Model H, 10 responses for each theme were summarized into one sentence using ChatGPT-4o. This summarization process aimed to retain the main points of Model H's original text while expressing them concisely.

3.1.1 The Future of Education

In the open-ended format, words representing abstract concepts such as "value," "role," "change," and "need" were characteristic. In contrast, the prompt engineering format featured words related to specific situation analysis and opportunities, such as "global," "potential," "current," "development," and "opportunity."

Model H's response summaries:

- Open-ended format: "Education needs to evolve into flexible and personalized learning models to respond to rapid technological innovation and social changes, while maintaining human values such as ethics, creativity, and critical thinking."

Table 1 Comparison of Frequently Occurring Words in Responses to Open-Ended and Prompt Engineering Questions Across Four Themes: Education, Corporate Communication, Meal Planning, and Movies.

theme	answer 1	answer 2
future education	value	global
	teacher	potential
	role	current
	system	development
	change	opportunity
	need	recommendation
communication	encourage	implementation
	help	effect
	foster	challenge
	work	platform
	create	term
	feedback	
	culture	
	organization	
	different	
	generational	
meal planning	wife	veggie
	help	skill
	helth	vegetable
	week	easy
	work	slow
	weekend	chicken
new movies	diverse	implementation
	develop	challenge
	explore	ar
	cultural	vr
	new	time
		real

- Prompt engineering format: "Focusing on the importance of technology, global challenges, and personalized learning in future education, it presents a flexible and comprehensive educational model and specific recommendations for educators and policymakers."

These results suggest that open-ended responses tend to focus on the intrinsic value and role of education and the need for change, while prompt engineering responses tend to focus on the international aspects of education, potential for development, and opportunity seeking based on current situation analysis.

3.1.2 Improving Internal Corporate Communication

In the open-ended format, words related to organizational culture and human relations, such as "encourage," "foster," "create," "feedback," and "culture," were characteristic. In contrast, the prompt engineering format featured words related to specific measures and effect measurement, such as "implementation," "effect," "challenge," "platform," and "term."

Model H's response summaries:

- Open-ended format: "Emphasizes the importance of interdepartmental collaboration, breaking down organizational silos, and fostering a culture of continuous learning and open communication to address common challenges such as communication errors, generational gaps, and departmental inconsistencies."
- Prompt engineering format: "These proposals emphasize strategic approaches to strengthen internal communication and promote information sharing across generations and departments."

These results suggest that open-ended responses tend to focus on fostering organizational culture, intergenerational communication, and the importance of feedback, while prompt engineering responses tend to focus on implementing specific measures, measuring effects, addressing challenges, and utilizing communication platforms.

3.1.3 Healthy Family Meal Planning

In the open-ended format, words related to family life and time management, such as "wife," "help," "health," "week," "work," and "weekend," were characteristic. In contrast, the prompt engineering format featured words related to specific ingredients and cooking methods, such as "veggie," "skill," "easy," "slow," and "chicken."

Model H's response summaries:

- Open-ended format: "Proposes a gradual and realistic approach to improving family healthy eating habits, emphasizing the importance of communication and teamwork."
- Prompt engineering format: "Suggests strategies and techniques for all family members to gradually learn how to prepare simple, nutritious meals, establishing sustainable eating habits within a busy daily routine."

These results suggest that open-ended responses tend to focus on family cooperation, health considerations, weekly planning, and balancing work and meals, while prompt engineering responses tend to focus on specific ingredients, cooking skills, and simple yet time-consuming cooking methods.

3.1.4 New Movies

In the open-ended format, words related to diversity and cultural aspects, such as "diverse," "develop," "explore," "cultural," and "new," were characteristic. In contrast, the prompt engineering format featured words related to new technology implementation and realism, such as "implementation," "challenge," "AR," "VR," "time," and "real."

Model H's response summaries:

- Open-ended format: "Explores new movie genres and technologies that dynamically adapt to viewers' emotions, choices, and cultural backgrounds, offering more immersive and interactive experiences."

- Prompt engineering format: "Proposes personalized, immersive interactive storytelling experiences utilizing AI technologies and VR/AR that adapt to viewers' emotions and choices."

These results suggest that open-ended responses tend to focus on diversity, exploration of cultural aspects, and the development of new filmmaking, while prompt engineering responses tend to focus on the implementation of new technologies (AR, VR), challenges, and real-time aspects.

3.1.5 Overall Trends

The frequent word analysis revealed characteristic trends for both open-ended and prompt engineering formats:

- **Open-ended format trends**:
  - More abstract and conceptual words appear (e.g., "value," "change," "culture," "diverse")
  - Responses often have a broader perspective and long-term considerations (e.g., "role," "foster," "develop")
  - Words often related to human factors and emotions (e.g., "need," "promote," "help")
- **Prompt engineering format trends**:
  - More concrete and practical words appear (e.g., "implementation," "skill," "platform")
  - Words often related to current analysis and short-term solutions (e.g., "current," "effect," "challenge")
  - Many words related to technology and methods (e.g., "AR," "VR," "vegetable," "time-consuming")

These trends suggest that the question format influences the focus and thought process of the responses. The open-ended format tends to encourage broader and more creative thinking, while the prompt engineering format tends to elicit more specific and actionable suggestions.

3.2 Analysis and Discussion of Evaluation Metrics Using Open-Source Tools

We quantitatively analyzed response characteristics using metrics such as text length (Length), Gunning Fog Index, Rix, and Measure of Textual Lexical Diversity (MTLD).

3.2.1 The Future of Education

- MTLD scores: Generally higher for open-ended questions, with statistically significant differences ($p<0.001$) observed in many models. This suggests that open-ended questions promote lexical diversity.
- Text length: Significantly longer for the prompt engineering format in all models ($p<0.001$). This indicates that structured questions elicit more detailed responses.

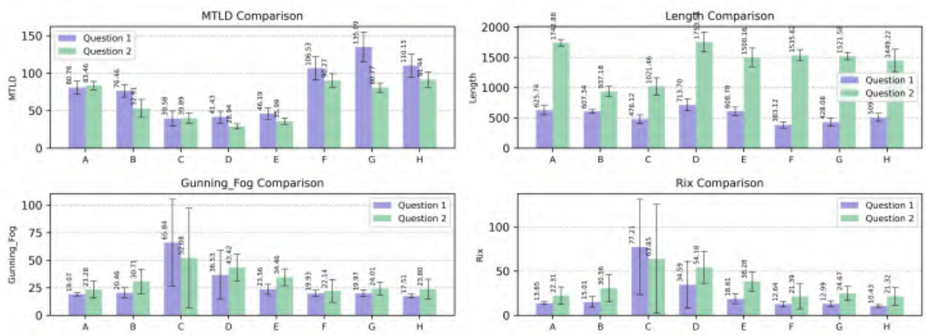

Fig. 1 Comparison of AI Model Performance on the "Future of Education" Theme.

Table 2 Open-source toolkit Metrics for Various Models, The Future of Education

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	80.8	8.8	83.5	5.5	76.5	8.3	52.9	12.0	39.6	10.0	39.9	6.7	41.4	7.8	28.9	3.5
Length	625.7	76.8	1742.9	52.6	607.3	29.0	937.2	80.5	476.1	71.8	1021.5	142.1	713.7	98.5	1753.5	162.8
Gunning	19.1	1.5	23.3	7.7	20.5	4.4	30.7	11.1	65.8	39.3	52.1	45.0	36.5	22.2	43.4	12.1
Rix	13.6	1.6	22.3	9.9	15.0	5.9	30.6	15.2	77.2	53.9	63.9	61.6	34.6	26.4	54.2	17.9

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	46.2	7.7	36.0	4.1	106.5	15.7	90.3	8.9	135.1	19.6	80.8	6.1	110.2	15.3	91.4	10.0
Length	608.8	68.3	1500.2	151.4	383.1	50.4	1535.6	92.5	428.1	67.3	1521.6	61.4	509.0	65.3	1449.2	184.0
Gunning	23.6	4.7	34.5	7.5	19.9	2.8	22.1	10.4	20.0	3.1	24.0	5.8	17.5	1.6	23.8	9.0
Rix	18.6	5.5	38.3	10.8	12.6	2.9	21.4	14.0	13.0	3.2	24.7	8.1	10.4	1.5	21.3	10.1

Table 3 Statistical analysis results for Open-source toolkit metrics across different models on the "Future of Education" Theme.Note: p-values are indicated as follows: p < 0.001 (***), p < 0.01 (**), p < 0.05 (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
MTLD	-1.8	0.07	11.4	***	-0.2	0.85	10.3	***	8.3	***	6.4	***	18.7	***	7.3	***
Length	-84.8	***	-27.3	***	-24.2	***	-38.6	***	-37.9	***	-77.4	***	-84.9	***	-34.1	***
Gunning	-3.8	***	-6.1	***	1.6	0.11	-1.9	0.06	-8.7	***	-1.4	0.15	-4.3	***	-4.9	***
Rix	-6.1	***	-6.8	***	1.2	0.25	-4.3	***	-11.5	***	-4.3	***	-9.5	***	-7.5	***

- Gunning Fog Index and Rix scores: Tended to be significantly higher for the prompt engineering format in many models. This suggests that the prompt engineering format tends to generate responses containing more complex and specialized expressions.

See **Fig.** 1 and **Tables** 2 and 3.

3.2.2 Improving Internal Corporate Communication

- MTLD scores: Significantly higher for the open-ended format in all models except Model H ($p<0.001$, Model C: $p<0.01$).

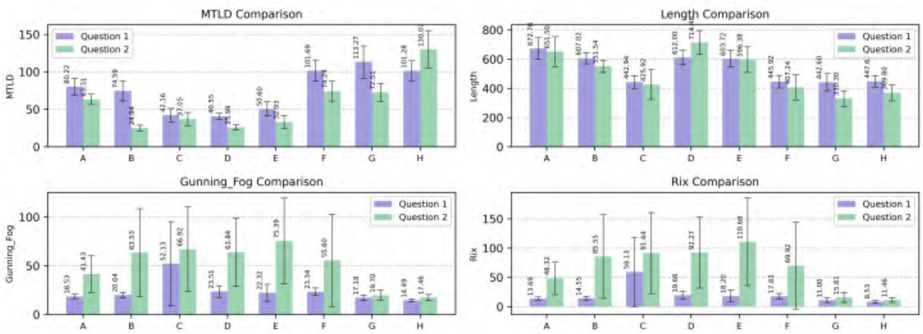

Fig. 2 Comparison of AI Model Performance on the "Improved communication within the company" Theme.

Table 4 Open-source toolkit Metrics for Various Models, Improved communication within the company

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	80.2	11.3	63.3	6.9	74.6	13.3	24.9	3.9	42.2	9.0	37.0	8.8	40.6	4.2	26.0	3.4
Length	672.8	74.3	651.5	103.2	607.0	39.0	551.5	42.7	442.9	44.5	425.9	103.1	612.0	49.2	714.5	82.2
Gunning	18.5	2.3	41.4	18.9	20.0	2.7	63.5	45.1	52.1	43.2	66.9	43.5	23.5	6.0	63.8	35.1
Rix	13.7	2.9	48.3	27.7	14.5	3.0	85.6	71.4	59.1	58.8	91.4	68.9	19.7	7.2	92.3	60.6

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	50.6	9.4	32.9	8.8	101.7	14.1	74.2	13.2	113.3	21.5	72.5	12.1	101.3	13.4	130.1	25.0
Length	603.7	58.1	596.4	88.0	445.9	43.6	407.2	86.3	442.6	57.7	330.3	53.8	447.6	39.5	369.8	55.0
Gunning	22.3	8.6	75.4	44.0	23.3	3.7	55.6	47.4	17.2	2.9	19.7	5.2	14.5	1.7	17.5	3.1
Rix	18.2	10.3	110.7	74.8	17.8	4.7	69.9	73.8	11.0	3.7	15.8	8.4	8.5	1.9	11.5	3.9

Table 5 Statistical analysis results for Open-source toolkit metrics across different models on the "Improving communication within companies" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
MTLD	9	***	25.3	***	2.9	**	19.2	***	9.7	***	10	***	11.7	***	-7.2	***
Length	1.2	0.24	6.8	***	1.1	0.29	-7.6	***	0.5	0.62	2.8	**	10.1	***	8.1	***
Gunning	-8.5	***	-6.8	***	-1.7	0.09	-8	***	-8.4	***	-4.8	***	-3	**	-6	***
Rix	-8.8	***	-7	***	-2.5	*	-8.4	***	-8.7	***	-5	***	-3.7	***	-4.8	***

- Text length: Showed different trends depending on the model, with some models showing significantly longer responses for open-ended questions and others showing no significant difference.
- Gunning Fog Index and Rix scores: Significantly higher for the prompt engineering format in most models. This suggests that the prompt engineering format tends to generate responses containing more complex and specialized content.

See **Fig.** 2 and **Tables** 4 and 5.

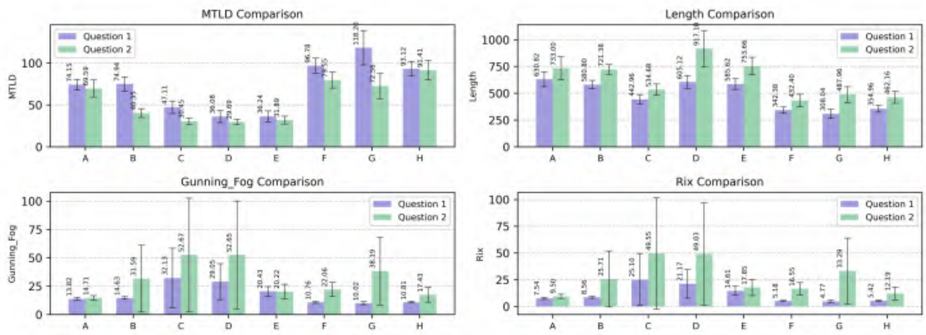

Fig. 3 Comparison of AI Model Performance on the "Healthy family meal planning" Theme.

Table 6 Open-source toolkit Metrics for Various Models, Healthy family meal planning

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	74.2	6.5	69.6	10.2	74.9	8.3	40.4	5.3	47.1	7.5	30.4	3.8	36.1	7.2	29.7	3.4
Length	630.8	69.3	733.0	108.1	580.8	39.2	721.4	45.2	443.0	45.0	534.7	52.6	605.1	59.5	917.1	168.7
Gunning	13.8	1.3	14.7	2.0	14.6	1.1	31.6	29.4	32.1	26.4	52.7	49.9	29.0	15.8	52.6	47.5
Rix	7.5	1.0	9.5	1.9	8.6	0.9	25.7	26.2	25.1	24.4	49.6	51.9	21.2	13.4	49.0	47.8

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	36.2	6.5	31.9	4.6	96.8	8.9	79.5	9.5	118.2	20.6	72.6	14.9	93.1	8.8	91.4	11.5
Length	585.6	50.4	753.7	81.4	342.4	29.7	432.4	58.1	308.0	42.7	488.0	73.4	355.0	32.0	462.2	56.2
Gunning	20.4	4.3	20.2	6.5	10.8	0.9	22.1	6.3	10.0	1.4	38.2	29.8	10.8	0.8	17.4	6.6
Rix	14.6	4.2	17.8	7.6	5.2	0.7	16.6	5.8	4.8	1.0	33.3	30.7	5.4	0.6	12.2	6.1

3.2.3 Healthy Family Meal Planning

- MTLD scores: Significantly higher for the open-ended format in all models except Model H ($p<0.001$, Model A: $p<0.01$). This suggests that the open-ended format encourages the generation of meal planning suggestions using a diverse range of vocabulary.
- Text length: Significantly longer for the prompt engineering format in all models ($p<0.001$). This indicates that the prompt engineering format elicits more detailed meal planning suggestions.
- Gunning Fog Index and Rix scores: Significantly higher for the prompt engineering format in most models ($p<0.001$ or $p<0.01$). This suggests that the prompt engineering format tends to generate more professional and specific meal planning suggestions.

See **Fig.** 3 and **Tables** 6 and 7

3.2.4 New Movies

- MTLD scores: Significantly higher for the open-ended format in all models except Model H ($p<0.001$, Model A: $p<0.05$). This suggests that the open-ended format promotes a broader discussion about movies.

Table 7 Statistical analysis results for Open-source toolkit metrics across different models on the "Healthy family meal planning" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
MTLD	2.7	**	24.9	***	14	***	5.7	***	3.8	***	9.4	***	12.7	***	0.8	0.41
Length	-5.6	***	-16.6	***	-9.4	***	-12.3	***	-12.4	***	-9.8	***	-15	***	-11.7	***
Gunning_Fog	-2.7	**	-4.1	***	-2.6	*	-3.3	**	0.2	0.85	-12.5	***	-6.7	***	-7.1	***
Rix	-6.4	***	-4.6	***	-3	**	-4	***	-2.6	**	-13.9	***	-6.6	***	-7.9	***

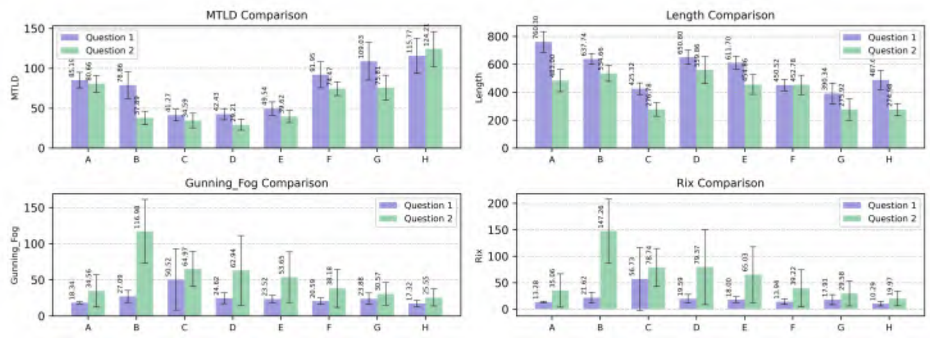

Fig. 4 Comparison of AI Model Performance on the "New Move" Theme.

Table 8 Open-source toolkit Metrics for Various Models, New Movie

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	85.2	9.8	80.7	10.4	78.9	16.8	37.9	8.4	41.3	7.2	34.6	9.1	42.4	7.0	29.2	6.8
Length	760.3	74.1	483.0	79.1	637.7	39.2	534.7	56.7	425.3	42.0	276.8	49.0	650.8	50.8	559.9	97.0
Gunning	18.3	1.4	34.6	22.2	27.1	8.2	117.0	44.0	50.5	42.2	65.0	24.4	24.6	7.4	62.9	48.3
Rix	13.3	1.6	35.1	31.6	21.6	10.1	147.3	60.4	56.7	59.3	78.7	35.8	19.6	8.3	79.4	71.1

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
MTLD	49.5	8.4	39.6	7.6	92.0	16.3	74.5	8.6	109.0	23.8	75.8	15.6	115.8	22.0	124.2	21.9
Length	611.7	45.6	454.9	70.7	450.5	42.0	452.8	69.4	390.3	74.8	275.9	77.1	487.6	69.1	275.0	43.8
Gunning	23.5	4.9	53.7	34.8	20.6	4.3	38.2	26.2	23.9	7.9	30.6	15.4	17.3	4.2	25.6	11.8
Rix	18.0	5.6	65.0	52.7	13.9	5.4	39.2	35.1	17.9	9.5	29.6	23.7	10.3	4.5	20.0	14.0

- Text length: Significantly longer for the open-ended format in all models except Model F ($p<0.001$). This suggests that the open-ended format encourages more extensive discussion about movies.
- Gunning Fog Index and Rix scores: Significantly higher for the prompt engineering format in all models ($p<0.001$, some $p<0.01$ or $p<0.05$). This suggests that the prompt engineering format tends to generate more specialized and specific responses about movie concepts and techniques.

See **Fig.** 4 and **Tables** 8 and 9.

Table 9 Statistical analysis results for Open-source toolkit metrics across different models on the "New films" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
MTLD	2.2	*	15.4	***	4.1	***	9.6	***	6.2	***	6.7	***	8.2	***	-1.9	0.06
Length	18.1	***	10.6	***	16.3	***	5.9	***	13.2	***	-0.2	0.84	7.5	***	18.4	***
Gunning_Fog	-5.1	***	-14.2	***	-2.1	*	-5.5	***	-6.1	***	-4.7	***	-2.7	**	-4.7	***
Rix	-4.9	***	-14.5	***	-2.2	*	-5.9	***	-6.3	***	-5.0	***	-3.2	**	-4.7	***

Table 10 Statistical analysis results for various metrics across different models on the "Future of Education" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
creativity	22.2	***	11.4	***	10.3	***	16.4	***	11.6	***	22.0	***	18.1	***	13.9	***
practicality	25.3	***	15.6	***	20.0	***	19.0	***	21.6	***	25.9	***	26.0	***	17.0	***
specificity	-18.3	***	-4.0	***	-7.2	***	-11.5	***	-8.9	***	-10.7	***	-16.5	***	-13.7	***
dialogue	-0.7	0.46	-4.4	***	4.8	***	7.7	***	6.7	***	-3.9	***	1.2	0.25	-3.7	***
thought	-4.1	***	4.8	***	1.6	0.11	3.5	***	0.2	0.85	2.9	**	1.8	0.07	0.7	0.49
complexity	-5.0	***	2.3	*	6.2	***	1.4	0.17	-3.0	**	-1.2	0.23	1.7	0.09	1.4	0.18
technicality	-10.8	***	-4.4	***	0.3	0.75	-1.4	0.15	-4.4	***	-10.5	***	-9.4	***	-8.3	***
diversity	46.6	***	32.1	***	29.1	***	33.6	***	36.6	***	72.1	***	53.6	***	34.6	***
coherence	-8.6	***	8.1	***	-2.2	*	3.9	***	1.6	0.10	5.0	***	5.2	***	0.4	0.66
readability	5.4	***	2.2	*	2.9	**	12.2	***	17.1	***	5.3	***	11.7	***	4.6	***
named	-24.0	***	-5.7	***	-4.9	***	-8.7	***	-8.6	***	-13.8	***	-14.0	***	-13.6	***
lexical	12.3	***	11.6	***	-2.9	**	-3.0	**	-3.9	***	9.7	***	10.9	***	11.4	***
avg. word	-0.7	0.48	-4.3	***	-3.1	**	-15.8	***	-19.3	***	-7.8	***	-12.0	***	-9.2	***
type token	50.0	***	25.0	***	27.9	***	31.5	***	31.6	***	53.5	***	52.5	***	33.2	***
dependency	-8.4	***	-2.5	*	-5.2	***	-7.3	***	-7.2	***	-10.3	***	-14.3	***	-14.5	***
passive	-0.7	0.49	-0.6	0.52	-1.7	0.09	1.4	0.16	-6.8	***	-0.9	0.35	-2.6	*	0.3	0.77

3.3 Analysis of Original Evaluation Metrics and Statistical Testing

We conducted a multifaceted analysis of AI model responses using 16 original evaluation metrics, including creativity, practicality, specificity, and naturalness of dialogue. All scores are shown in Appendix 22,23,24,25,26,27,28,29.

3.3.1 The Future of Education

The open-ended format was statistically significantly superior to the prompt engineering format in terms of creativity, practicality, and diversity (see Figure 5, Table 10). In particular, Models A, F, and G showed very high t-values for these indicators, demonstrating superior ability to generate creative, practical, and diverse responses in the field of education. On the other hand, in terms of specificity and technicality, the prompt engineering format was dominant in almost all models.

3.3.2 Improving Internal Corporate Communication

In terms of creativity and practicality, the open-ended format was slightly dominant in most models, but only some models showed statistically significant differences (see

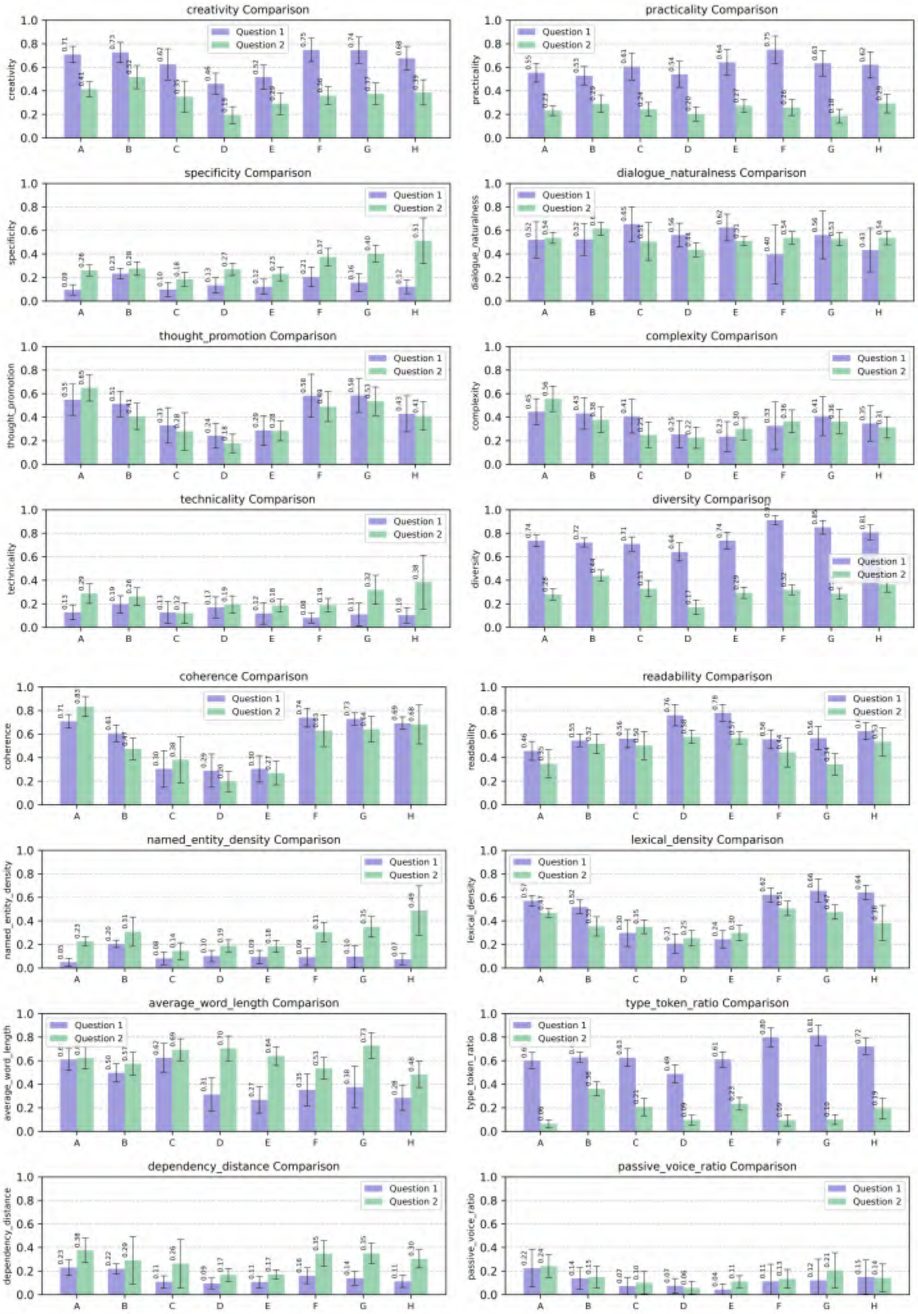

Fig. 5 Comparison of AI Model Performance on the "Future of Education" Theme.

Table 11 Statistical analysis results for various metrics across different models on the "Improving communication within companies" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
creativity	1.5	0.14	3.7	***	-1.8	0.07	11.6	***	2.4	*	0.8	0.45	4.1	***	-0.4	0.67
practicality	6.6	***	0.8	0.44	-0.7	0.47	6.2	***	5.1	***	1.7	0.10	3.6	***	0.0	0.96
specificity	-3.5	***	-3.7	***	-2.6	*	-6.6	***	-10.3	***	-1.3	0.21	-6.5	***	-6.3	***
dialogue	-7.0	***	-7.8	***	-3.2	**	4.6	***	6.2	***	-3.0	**	-4.8	***	-8.2	***
thought	-4.9	***	9.4	***	-1.1	0.29	10.5	***	4.8	***	2.5	*	7.0	***	4.0	***
complexity	-6.1	***	5.7	***	1.6	0.11	9.0	***	9.4	***	-3.2	**	-0.3	0.76	0.2	0.83
technicality	-1.5	0.13	-6.1	***	-0.3	0.76	-5.4	***	-3.3	**	-10.2	***	-3.9	***	-8.9	***
diversity	2.0	*	6.5	***	1.4	0.17	15.2	***	16.1	***	8.6	***	7.2	***	-5.6	***
coherence	-5.7	***	6.9	***	-3.8	***	8.8	***	-0.7	0.47	-0.4	0.67	5.7	***	-1.5	0.14
readability	11.3	***	11.0	***	8.3	***	8.9	***	14.3	***	9.1	***	5.5	***	0.5	0.63
named	-0.9	0.36	-0.8	0.43	-0.4	0.69	-7.5	***	-7.6	***	-10.2	***	-12.7	***	-3.2	**
lexical	-3.2	**	5.7	***	-2.8	**	0.7	0.46	0.0	0.99	1.8	0.08	3.6	***	-4.2	***
avg. word	-18.2	***	-23.0	***	-9.8	***	-14.0	***	-18.5	***	-17.5	***	-13.5	***	-0.1	0.89
type token	0.9	0.40	-6.2	***	1.3	0.20	15.3	***	7.8	***	2.5	*	0.6	0.56	-8.8	***
dependency	-8.4	***	-7.1	***	-4.8	***	2.0	*	1.4	0.15	-3.1	**	-5.5	***	0.1	0.93
passive	3.9	***	6.1	***	1.9	0.06	1.0	0.31	5.4	***	-0.1	0.90	5.4	***	3.2	**

Table 12 Statistical analysis results for various metrics across different models on the "Healthy family meal planning" Theme.Note: p-values are indicated as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
creativity	-1.9	0.06	6.4	***	-4.2	***	4.7	***	5.2	***	-0.4	0.66	-2.8	**	4.2	***
practicality	19.8	***	20.7	***	15.3	***	18.7	***	23.1	***	12.4	***	18.3	***	15.4	***
specificity	8.1	***	-0.5	0.64	-1.4	0.17	-2.2	*	-5.9	***	-1.7	0.10	7.6	***	-5.5	***
dialogue	-12.6	***	-18.9	***	-8.0	***	-3.5	***	-2.8	**	-23.1	***	-7.4	***	-12.3	***
thought	1.6	0.11	8.8	***	2.4	*	3.5	***	4.2	***	-0.2	0.83	-0.8	0.40	7.0	***
complexity	7.0	***	2.8	**	4.7	***	4.2	***	1.6	0.11	-5.0	***	0.1	0.95	10.5	***
technicality	-14.6	***	-15.5	***	-5.0	***	-4.8	***	-4.4	***	-7.6	***	-4.8	***	-2.2	*
diversity	1.1	0.26	6.5	***	-3.8	***	15.1	***	5.3	***	6.6	***	13.1	***	9.1	***
coherence	-8.8	***	5.5	***	0.0	0.96	-0.9	0.39	1.6	0.11	-0.1	0.89	-7.0	***	-2.6	*
readability	6.5	***	0.4	0.66	2.7	**	9.4	***	4.5	***	10.5	***	11.4	***	4.3	***
named	6.0	***	-6.0	***	-2.0	*	-4.5	***	-8.1	***	-1.1	0.26	9.9	***	-2.9	**
lexical	1.7	0.10	4.8	***	2.9	**	-1.0	0.32	0.4	0.67	0.3	0.79	0.0	0.96	-2.1	*
avg. word	-6.3	***	-6.4	***	-4.5	***	-8.6	***	-7.1	***	-13.0	***	-16.2	***	-3.9	***
type token	-0.9	0.38	3.0	**	-6.7	***	9.2	***	-1.4	0.16	2.6	*	9.6	***	8.2	***
dependency	-9.8	***	-5.8	***	-4.9	***	-7.0	***	-3.1	**	-11.6	***	-5.4	***	-5.4	***
passive	-2.5	*	2.7	**	0.3	0.74	-3.2	**	-4.5	***	-4.7	***	1.1	0.27	-2.5	*

Figure 6, Table 11). In terms of diversity, the open-ended format was significantly superior in all models except Model H ($p<0.001$). On the other hand, in terms of dialogue naturalness, the prompt engineering format was significantly superior in many models.

3.3.3 Healthy Family Meal Planning

In terms of practicality, the open-ended format was statistically significantly superior in all models (see Figure 7, Table 12). For creativity, the dominant format differed depending on the model. In terms of dialogue naturalness, the prompt engineering format was significantly superior for all models.

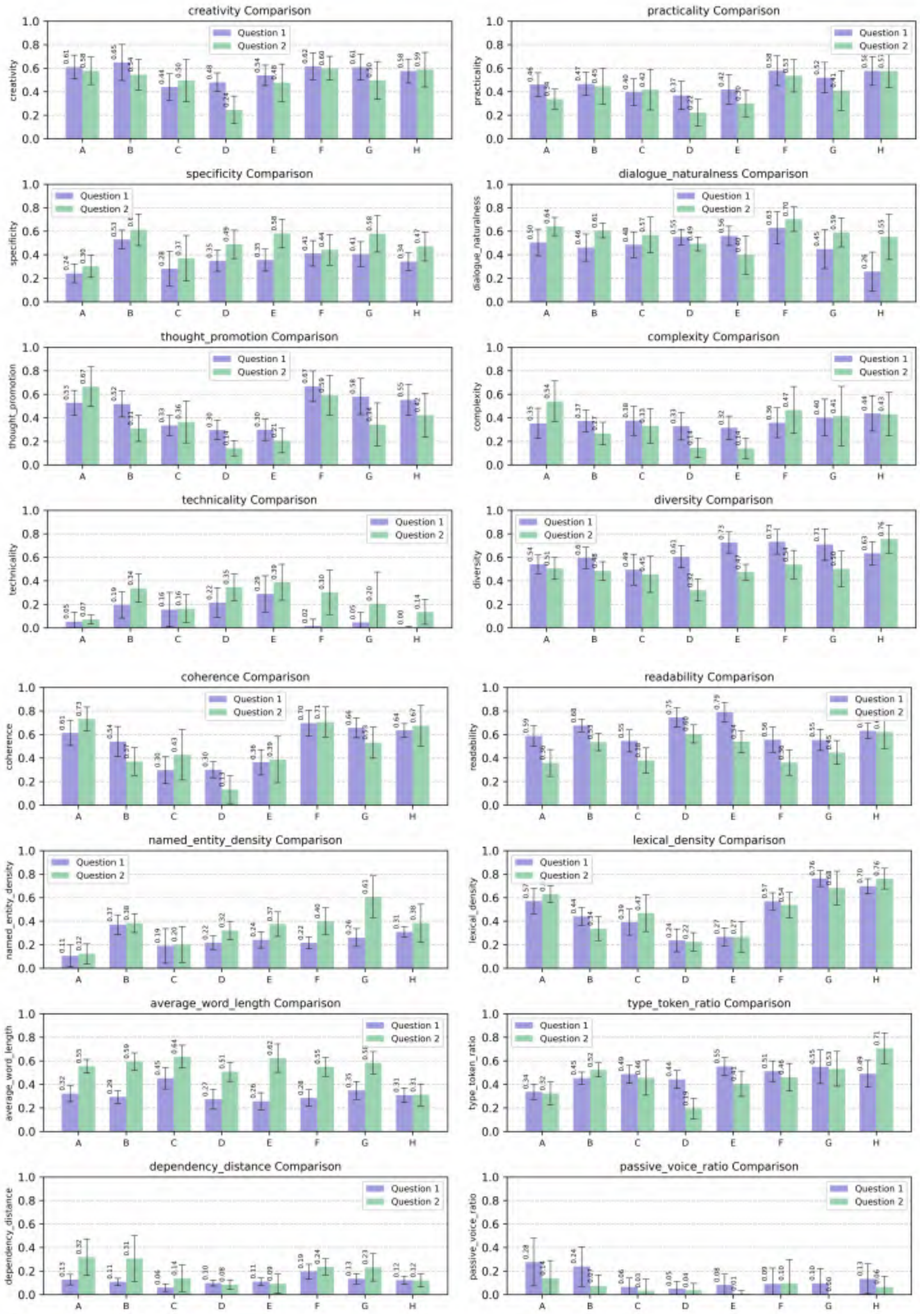

Fig. 6 Comparison of AI Model Performance on the "Improving communication within companies" Theme.

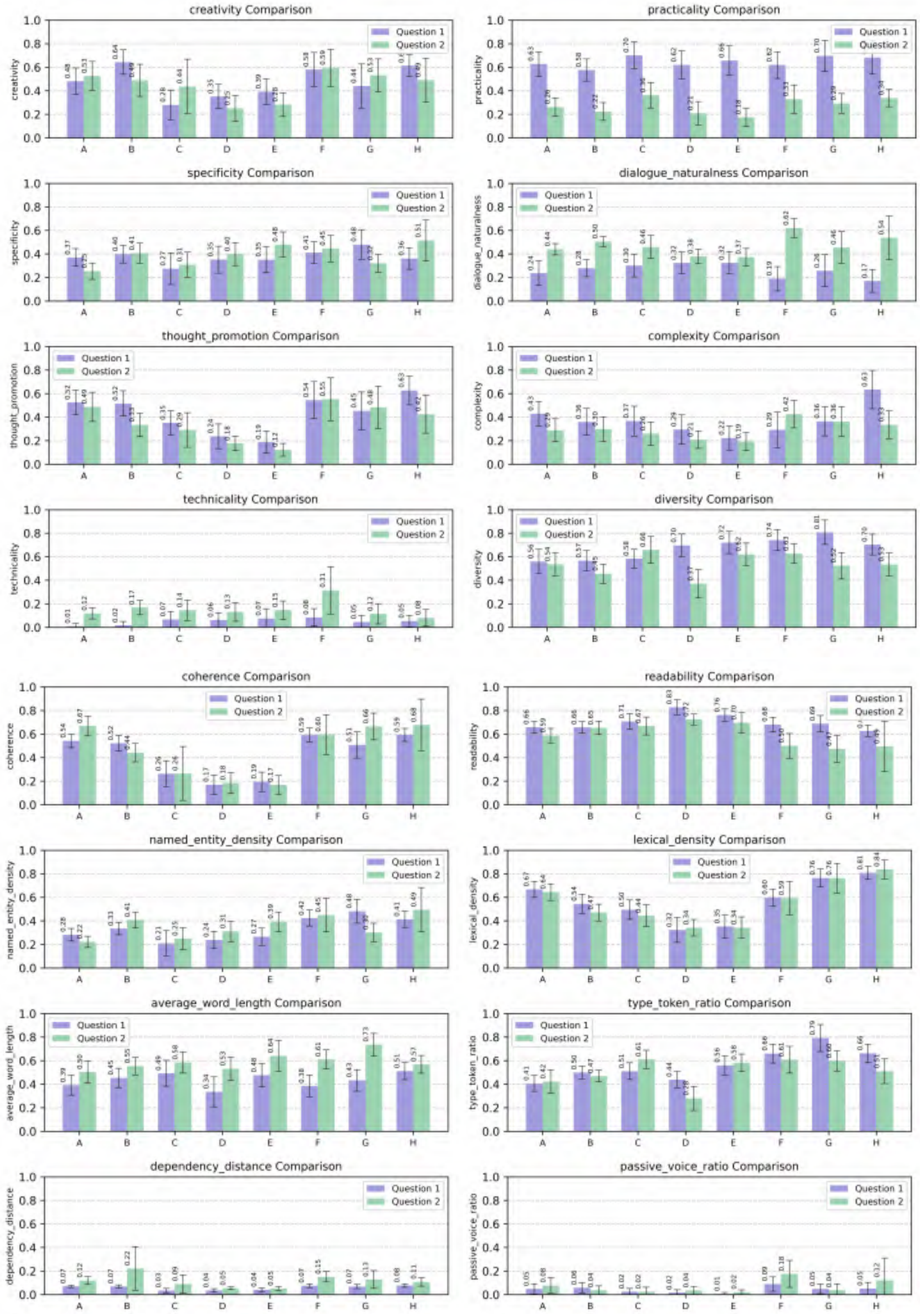

Fig. 7 Comparison of AI Model Performance on the "Healthy family meal planning" Theme.

Table 13 Statistical analysis results for various metrics across different models on the "New films" Theme.Note: p-values are indicated as follows: p < 0.001 (***), p < 0.01 (**), p < 0.05 (*). Values greater than or equal to 0.05 are displayed with two decimal places.

model	A		B		C		D		E		F		G		H	
metric	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
creativity	-3.5	***	-0.5	0.63	-2.6	*	6.2	***	-1.0	0.31	0.6	0.58	4.8	***	0.7	0.49
practicality	-5.9	***	-7.2	***	-2.5	*	3.9	***	3.5	***	6.8	***	9.5	***	1.5	0.14
specificity	-2.6	**	5.3	***	3.9	***	-6.7	***	-12.7	***	1.3	0.20	-9.0	***	-4.6	***
dialogue	-6.7	***	-3.2	**	-2.4	*	0.2	0.87	2.4	*	-7.5	***	-1.3	0.21	-12.7	***
thought	-4.8	***	1.6	0.11	1.7	0.10	4.0	***	1.4	0.17	-1.4	0.17	8.6	***	0.3	0.77
complexity	-11.2	***	-0.1	0.89	0.2	0.85	0.6	0.55	3.3	**	-7.6	***	0.8	0.44	-2.3	*
technicality	-7.6	***	-4.8	***	3.4	***	-5.0	***	-6.8	***	-2.7	**	-8.2	***	-4.5	***
diversity	-9.1	***	4.3	***	-5.9	***	7.7	***	7.3	***	8.7	***	2.6	*	0.7	0.47
coherence	-4.0	***	-2.9	**	-0.1	0.94	6.4	***	-0.5	0.64	1.2	0.24	9.2	***	-1.2	0.25
readability	7.5	***	10.1	***	1.0	0.31	11.5	***	11.9	***	13.6	***	1.7	0.10	6.0	***
named	-7.6	***	1.6	0.10	1.0	0.34	-4.8	***	-9.6	***	-1.1	0.29	-12.0	***	-8.0	***
lexical	2.0	0.05	1.9	0.07	-0.6	0.52	0.9	0.37	0.1	0.92	1.3	0.20	9.5	***	-5.5	***
avg. word	-8.4	***	-10.8	***	-1.4	0.17	-13.6	***	-12.9	***	-13.9	***	-9.4	***	-3.6	***
type token	-12.5	***	-2.6	**	-10.6	***	2.9	**	-1.4	0.16	7.6	***	-0.1	0.96	-7.3	***
dependency	-5.0	***	-6.6	***	-3.6	***	-0.1	0.95	3.1	**	-3.5	***	0.3	0.79	-3.8	***
passive	-0.2	0.87	5.6	***	0.0	0.99	4.8	***	3.1	**	-0.4	0.71	5.7	***	-0.5	0.59

3.3.4 New Movies

For creativity and practicality, the dominant format differed depending on the model (see Figure 8, Table 13). In terms of specificity and dialogue naturalness, the prompt engineering format was significantly dominant for most models. For diversity, there were models where the open-ended format was dominant and others where the prompt engineering format was dominant.

3.3.5 Overall Trends

Overall, the open-ended format tended to be dominant in terms of creativity, practicality, diversity, and readability, while the prompt engineering format tended to be dominant in terms of specificity, dialogue naturalness, and technicality. However, the optimal format differed depending on the theme and model, and the results were not generalizable.

3.4 Interaction between AI Model Characteristics and Question Format

The results showed that the interaction between AI model characteristics and question format significantly impacts the quality and effectiveness of the dialogue.

3.4.1 Comparison between Models

Models F, G, and H scored highly on many metrics and performed particularly well on the themes of education and internal communication. These models showed high adaptability to both question formats and consistently performed well.

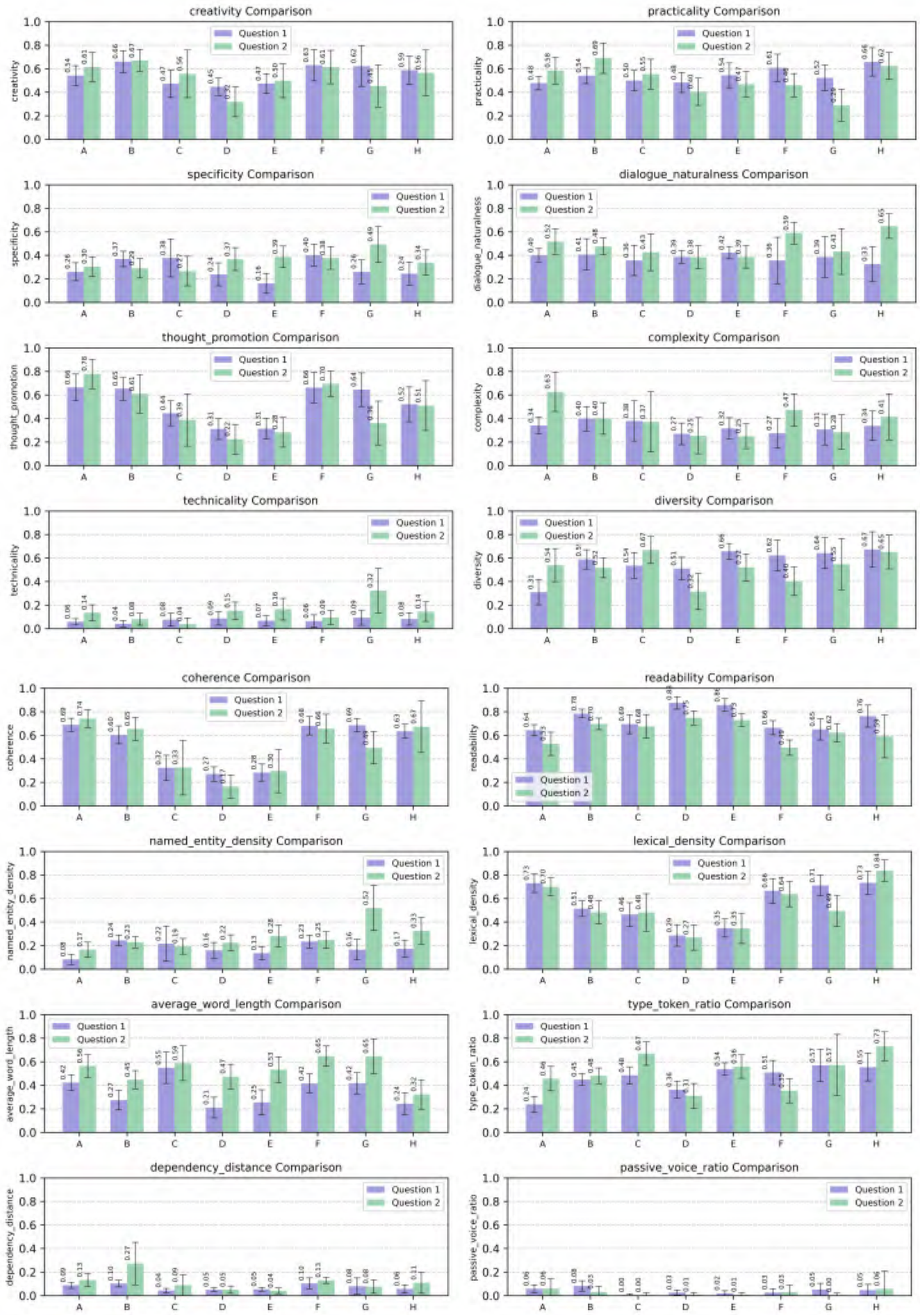

Fig. 8 Comparison of AI Model Performance on the "New films" Theme.

On the other hand, the effects of the question format were more pronounced in other models. For example, in the education theme, Model A showed a very high t-value of 22.2 ($p<0.001$) for the creativity metric in the open-ended format, while it dropped significantly to -18.3 ($p<0.001$) in the prompt engineering format.

3.4.2 Theme Dependency

The effectiveness of question types varied greatly depending on the theme. For instance, the superiority of the open-ended format was most pronounced in the education theme. On the other hand, in the movie theme, the effects of question formats varied widely across models, showing no consistent trend.

3.4.3 Relationship between Evaluation Metrics

The metrics for creativity, practicality, and diversity often showed similar trends, with open-ended questions being dominant. On the other hand, the indicators of specificity, naturalness of dialogue, and technicality also often showed similar trends, with prompt engineering formats being dominant.

These results suggest that the knowledge structure and language processing algorithms of AI models influence how questions are interpreted and responses are generated.

4 Discussion

This study compared and analyzed the effects of question formats in dialogue with generative AI. We examined the effects of open-ended questions and prompt engineering questions using eight AI models across four themes, analyzing them from multiple perspectives using 20 evaluation metrics. The results provide important implications for the design and application of computational intelligence systems, particularly conversational AI.

4.1 Multifaceted Effects of Question Formats

The results demonstrate that question formats have a multifaceted and complex impact on AI model responses. Open-ended questions showed advantages in terms of creativity, diversity, and thought promotion, while prompt engineering formats were effective in terms of concreteness and naturalness of dialogue.

This finding has significant implications for the design of human-AI interactions in computational intelligence systems. For example, the fact that open-ended questions showed statistically significantly higher scores on the creativity index in the education theme (e.g., t-value of 22.2, $p<0.001$ for Model A) suggests the possibility of actively utilizing open-ended questions to promote creative thinking in the design of educational AI systems.

On the other hand, the superiority of the prompt engineering format in terms of dialogue naturalness in the corporate communication theme (e.g., t-value of -8.2, $p<0.001$ for Model H) provides insights that can be applied to the design of business-oriented AI assistants. This aligns with Bozkurt's (2024) assertion, which positions prompt engineering as a new digital competency and emphasizes its importance [4]. While Bozkurt does not present empirical research results, our findings support his claims.

4.2 Theme Dependency and AI Model Characteristics

We also found that the effectiveness of question formats varied greatly depending on the theme. For instance, while open-ended questions showed clear superiority in the education theme, the effects of question formats varied widely across models in the movie theme. This finding suggests the need for adaptive dialogue strategies in the design of computational intelligence systems, depending on the domain and topic being addressed.

Furthermore, we discovered that the interaction between AI model characteristics and question formats significantly impacts the quality and effectiveness of dialogue. For example, some models demonstrated high adaptability to both question formats, consistently performing well. This result suggests the effectiveness of the prompt pattern catalog proposed by White et al. (2023) [6]. Additionally, our findings indicate that prompt engineering is a context-dependent and complex process, emphasizing the need for flexible and adaptive approaches based on AI models and dialogue themes.

Notably, we observed an unexpected phenomenon where the effects of question formats were reversed in some models. For instance, in the movie theme, the prompt engineering format showed superiority in the creativity index for Model A (t-value -3.5, $p<0.001$), while open-ended questions were superior for Model D (t-value 6.2, $p<0.001$). This result suggests that the internal structure and training data of AI models may significantly influence the effectiveness of question formats, warranting further investigation in future research.

4.3 Theoretical and Practical Contributions to Computational Intelligence Systems

The results of this study make significant contributions to the field of computational intelligence systems, particularly in natural language processing and dialogue systems. Firstly, by quantitatively demonstrating the impact of question formats on AI response characteristics, we provide a theoretical foundation for developing more effective human-AI dialogue models. This empirically supports the potential of prompt engineering in large language models, as proposed by Chen et al. (2023) [11]. While Chen et al. pointed out the important role of prompt engineering in leveraging the capabilities of large language models, our study concretizes this claim from the perspective of the effectiveness of different question formats.

From a practical standpoint, the insights from this study can be directly applied to the design of next-generation conversational AI systems. For example, it is possible to optimize questioning strategies according to the purpose, such as prioritizing open-ended questions to promote creative thinking in educational support AI, and using prompt engineering formats to elicit specific information in business-oriented AI assistants.

Moreover, our findings on the interaction between AI model characteristics and question formats provide guidelines for selecting optimal AI models based on tasks and situations. This is crucial knowledge that can lead to efficient use of computational resources and improved user experience.

4.4 Limitations and Future Research Directions

This study has several limitations. First, as the experiments were conducted only in English, caution is needed regarding generalizability to multilingual environments. There may also be biases in the selected themes and AI models. These limitations, considering the complexity and diversity of AI prompt engineering pointed out by Oppenlaender et al., suggest the need for further research [18]. While Oppenlaender et al. emphasize the increasing importance of prompt engineering as a new digital competency, our study underscores the need for further exploration in this field.

Based on these limitations, we propose the following future research directions:

1. Verification in multilingual and multicultural environments: Investigating the effects of question formats in different languages and cultural backgrounds can yield more universal insights.
2. Long-term dialogue analysis: While this study analyzed short-term dialogues, examining the effects of question formats in long-term dialogues could provide more practical insights.
3. Improvement of AI models: Analyzing the characteristics of models that showed high adaptability to question formats can lead to the development of more flexible and effective AI models.
4. User interface design: Designing and evaluating user interfaces that can maximize the effects of question formats is an important research topic.

4.5 Ethical Considerations and Social Impact

As AI dialogue becomes increasingly pervasive in daily life, ethical considerations are becoming increasingly important. The results of this study show that AI systems generate different responses depending on the question format, suggesting that AI decisions and recommendations may vary greatly depending on how questions are posed.

This finding emphasizes the importance of AI literacy education. Understanding the impact of question formats and learning effective communication methods with AI can lead to appropriate use of AI and reduction of potential biases. This relates to the challenges in students' perception of AI pointed out by Marrone et al. (2022) [19]. While Marrone et al.'s study showed that students do not fully recognize the value of everyday applications of AI, our research emphasizes the importance of understanding effective dialogue methods with AI, further supporting the need for AI literacy education.

Furthermore, AI system designers and developers have the responsibility to ensure consistency and fairness in AI responses to different question formats. This is crucial for enhancing the reliability and social acceptability of AI systems.

5 Conclusion

This study represents an important step towards optimizing human-AI dialogue in computational intelligence systems. By understanding and appropriately utilizing the influence of question formats, more effective and creative human-AI collaboration becomes possible. Future research is expected to further develop the insights gained here, leading to the development of more sophisticated dialogue AI systems adaptable to diverse situations and cultural backgrounds.

These findings will have a significant impact on both the theory and practice of computational intelligence systems, forming the foundation for more effective collaboration between AI and humans.

6 Acknowledgements

In this paper, we received assistance from Claude 3.5 Sonnet, an advanced language model, for simplifying complex sentences, proofreading the English manuscript, and structuring the explanation of experimental results and discussion. However, all content has been rigorously reviewed, verified, and edited by the authors, ensuring technical accuracy, and contextual appropriateness. This approach allowed us to leverage AI capabilities while maintaining the integrity and originality of our scientific contribution.

Statements and Declarations

Competing Interests

The author declares no competing interests.

Data Availability

The conversation data used in this study is available from the author upon reasonable request.

Author's Contribution

This is a single-author paper. The author is responsible for the study conception, design, data analysis, and manuscript writing.

Funding

This research received no external funding.

References

[1] Eapen, T., Finkenstadt, D.J., Folk, J., Venkataswamy, L.: How generative ai can augment human creativity. Harvard Business Review **101**(4) (2023)

[2] Rayan, J., Kanetkar, D., Gong, Y., Yang, Y., Palani, S., Xia, H., Dow, S.P.: Exploring the potential for generative ai-based conversational cues for real-time collaborative ideation. In: Proceedings of the 16th Conference on Creativity & Cognition, pp. 117–131 (2024)

[3] Yang, D.: Human-ai interaction in the age of large language models. In: Proceedings of the AAAI Symposium Series, vol. 3, pp. 66–67 (2024)

[4] Bozkurt, A.: Tell me your prompts and I will make them true: The alchemy of prompt engineering and generative AI. International Council for Open and Distance Education Oslo, Norway (2024). https://doi.org/10.55982/openpraxis.16.2.661.

[5] Korzyński, P., Mazurek, G., Krzypkowska, P., Kurasiński, A.: Artificial intelligence prompt engineering as a new digital competence: Analysis of generative ai technologies such as chatgpt. Entrepreneurial Business and Economics Review (2023) https://doi.org/10.15678/eber.2023.110302

[6] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT (2023). https://arxiv.org/abs/2302.11382

[7] White, J., Hays, S., Fu, Q., Spencer-Smith, J., Schmidt, D.C.: ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design (2023). https://arxiv.org/abs/2303.07839

[8] Wit, A.: 'when i'm retired...': Using topic modeling to analyze open-ended survey questions in the a broader mind longitudinal survey. (2021). https://doi.org/10.31235/osf.io/ng4m2 . https://api.semanticscholar.org/CorpusID:240821164

[9] Haq, I.U., Pifarré, M., Fraca, E.: Novelty evaluation using sentence embedding models in open-ended cocreative problem-solving. International Journal of Artificial Intelligence in Education (2024) https://doi.org/10.1007/s40593-024-00392-3

[10] Ecoffet, A., Clune, J., Lehman, J.: Open Questions in Creating Safe Open-ended AI: Tensions Between Control and Creativity (2020). https://arxiv.org/abs/2006.07495

[11] Chen, B., Zhang, Z., Langrené, N., Zhu, S.: Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv preprint arXiv:2310.14735 (2023)

[12] Boussioux, L., Lane, J.N., Zhang, M., Jacimovic, V., Lakhani, K.R.: The crowdless future? generative ai and creative problem solving. Working Paper 24-005, Harvard Business School Technology & Operations Mgt. Unit (July 2024). https://doi.org/10.2139/ssrn.4533642 . https://ssrn.com/abstract=4533642

[13] Doshi, A.R., Hauser, O.P.: Generative ai enhances individual creativity but reduces the collective diversity of novel content. Science Advances **10**(28), 5290 (2024) https://doi.org/10.1126/sciadv.adn5290 https://www.science.org/doi/pdf/10.1126/sciadv.adn5290

[14] Perifanis, N.-A., Kitsios, F.: Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. Information **14**(2) (2023) https://doi.org/10.3390/info14020085

[15] Markowitz, D.M., Boyd, R.L., Blackburn, K.: From silicon to solutions: Ai's impending impact on research and discovery. Frontiers in Social Psychology **2** (2024) https://doi.org/10.3389/frsps.2024.1392128

[16] Anthropic: Claude prompting guide.md. Claude AI assistant. Unpublished document, Claude Example Project, as of August 18, 2024 (2024)

[17] Shimabucoro, L., Ruder, S., Kreutzer, J., Fadaee, M., Hooker, S.: LLM See, LLM Do: Guiding Data Generation to Target Non-Differentiable Objectives (2024). https://arxiv.org/abs/2407.01490

[18] Oppenlaender, J., Linder, R., Silvennoinen, J.: Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering (2024). https://arxiv.org/abs/2303.13534

[19] Marrone, R., Taddeo, V., Hill, G.: Creativity and artificial intelligence—a student perspective. Journal of Intelligence **10**(3), 65 (2022) https://doi.org/10.3390/jintelligence10030065

A Appendix

A.1 Questions

A.1.1 Future of Education

Open-ended Questioning Technique
As a high school teacher, I've noticed that recent technological advancements and rapid social changes are creating challenges that traditional educational models struggle to address.

For instance, the development of AI and robotics may significantly alter future job landscapes. Additionally, there's an urgent need to cultivate individuals capable of tackling global issues like climate change and political-economic fluctuations.

We must also consider the diversity of learners and the importance of mental health support, especially given the increase in such cases since the COVID-19 pandemic. The pandemic has also accelerated the use of ICT in education.

Beyond computers, developments in virtual reality and neuroscience may open new possibilities for education. While embracing these changes, I believe it's crucial to maintain core educational values such as fostering humanity, creativity, and critical thinking skills.

The importance of teaching manners, ethics, and integrity seems to be growing as well. These considerations lead me to question various aspects of our educational system: the relevance of physical school spaces (is gathering in classrooms necessary?), the changing role of teachers (with the availability of high-quality on-demand classes), and the validity of our grading systems (is there still value in closed-book exams?).

What do you think?

Prompt engineering style questions

As an experienced high school teacher and education futurist, please analyze the following aspects of education in light of recent technological advancements and social changes. For each point, provide a brief analysis of the current situation, potential future developments, and recommendations for educators and policymakers.

1. Job Market Transformation
   - Impact of AI and robotics on future employment
   - Skills needed for emerging industries
2. Global Challenges
   - Preparing students to address issues like climate change and political-economic fluctuations
   - Developing global citizenship and cross-cultural competencies
3. Learner Diversity and Mental Health
   - Strategies for inclusive education
   - Mental health support in schools, especially post-COVID
4. Technology in Education
   - Role of ICT, virtual reality, and neuroscience in learning
   - Balancing technology use with traditional teaching methods
5. Core Educational Values
   - Fostering humanity, creativity, and critical thinking
   - Teaching ethics, integrity, and digital citizenship
6. Educational Structures
   - Relevance of physical school spaces
   - Evolving role of teachers
   - Validity of current assessment methods

For each of these six areas:

1. Describe the current challenges and opportunities
2. Predict potential developments in the next 10-15 years
3. Suggest 2-3 actionable recommendations for educators and school administrators

After addressing these points, please:

1. Summarize the key trends that will likely shape the future of education
2. Identify any potential risks or unintended consequences of these changes
3. Propose an innovative educational model that addresses these challenges while maintaining core educational values

Throughout your analysis, please cite relevant research or case studies where appropriate. If you're uncertain about any predictions or recommendations, please acknowledge this uncertainty.

A.1.2 Improving Internal Corporate Communication

Open-ended Questioning Technique

In our company, the lack of communication between departments has become a major issue. In particular, there's a lack of smooth information sharing between the sales and development departments, leading to project delays and mismatches with customer needs. We've tried implementing regular joint meetings and introducing chat tools, but we haven't reached a fundamental solution.

Recently, an employee said, "If we were one living organism, information would flow more smoothly," and I feel there might be a hint there. But I'm not sure how to practically apply this idea.

Generational communication gaps are also an issue, with conflicts arising due to differences in values and working styles between veteran employees and younger staff.

We want to promote natural and organic communication that transcends the vertical organizational structure, but we feel limited by conventional methods. I think we need new perspectives and innovative approaches. What do you think?

Prompt engineering style questions

You are an expert in organizational development and improving internal communication. Please suggest ideas for improving our company's internal communication, considering the following points:

1. Objective: To resolve the lack of interdepartmental communication and achieve more effective and comprehensive information sharing.
2. Current challenges:
   (a) Lack of information sharing between sales and development departments
   (b) Project delays and mismatches with customer needs
   (c) Generational communication gaps
   (d) Information silos due to vertical organizational structure
3. Previous efforts:
   (a) Regular joint meetings
   (b) Introduction of chat tools
4. Elements to consider:
   (a) Utilization of latest communication technologies
   (b) Differences in values and work styles between generations
   (c) Organizational culture change
   (d) Promotion of natural and organic communication
5. Constraints:
   (a) Avoid major changes to existing organizational structure

    (b) Ensure privacy and information security
    (c) Consider implementation costs and operational efficiency
6. Expected outcomes:
    (a) Proposal of at least 5 specific improvement measures
    (b) Implementation methods and expected effects for each proposal
    (c) Anticipated challenges and countermeasures
7. Additional instructions:
    (a) Present proposals in bullet points clearly.
    (b) Add a concise explanation of about 150 characters for each proposal.
    (c) Consider the balance between innovative ideas and feasibility.
    (d) Distinguish between short-term implementable measures and long-term strategies.

Based on this task, please propose innovative and feasible strategies for improving our company's internal communication.

A.1.3 Healthy Family Meal Planning

Open-ended Questioning Technique
In our busy daily lives, we often rely on convenience food or eating out. However, my recent health check-up results weren't good, and these eating habits are also becoming expensive. I want to consider simple, delicious, and nutritionally balanced meals for my family's health.

The challenge is, I can't easily ask my wife to cook an extra meal, as she's already busy taking care of the kids and preparing three lunch boxes every day. I thought about waking up earlier to help, but I'm not good at cooking. Especially after my wife got angry at me for leaving the frying pan dirty, I've developed a mental block against cooking. Moreover, I come home late from work and want to sleep in when I can.

What do you think about this situation?

Prompt engineering style questions
You are an expert in nutrition and family meal planning. Please provide advice on creating simple, nutritious, and cost-effective meal plans for a busy family, considering the following points:

1. Objective: To develop easy-to-prepare, healthy meal options that fit into a busy lifestyle and improve overall family health.
2. Current challenges:
    (a) Reliance on convenience food and eating out
    (b) Poor health check-up results
    (c) Increasing food expenses
    (d) Limited time for meal preparation
    (e) Limited cooking skills of one family member
    (f) Mental block against cooking due to past experiences
3. Family situation:
    (a) Working parents with children
    (b) One parent prepares three lunch boxes daily
    (c) Late work hours for one parent

4. Constraints:
   (a) Minimal cooking time available
   (b) Need for simple recipes suitable for beginners
   (c) Budget considerations
5. Expected outcomes:
   (a) Proposal of at least 5 easy, nutritious meal ideas
   (b) Time-saving meal preparation strategies
   (c) Tips for overcoming mental blocks related to cooking
6. Additional instructions:
   (a) Present meal ideas and strategies in clear bullet points
   (b) Provide a brief explanation (about 50 words) for each suggestion
   (c) Include ideas for gradual skill improvement in cooking
   (d) Consider ways to involve all family members in meal preparation

Based on these requirements, please provide practical and innovative meal planning strategies that can improve this family's eating habits and overall health.

A.1.4 New Movie

Open-ended Questioning Technique
I'm in charge of developing new content at a film production company. Recently, I've felt that traditional movie genres and storytelling techniques are no longer fully meeting the needs of our increasingly diverse audience.

A creator recently said, "What if movies could directly experience the audience's emotions and thoughts?" This made me feel there might be new possibilities, but I'm struggling with how to actually realize this.

Also, with the development of AI and VR technologies, it's becoming possible to create new forms of movies that incorporate interactive elements. We're exploring new forms of audience-participatory entertainment that go beyond traditional linear storytelling.

Furthermore, as globalization progresses, developing universal storytelling techniques that resonate across cultural and language barriers is also a challenge.

We want to develop new movie genres and storytelling techniques that go beyond the traditional concept of film, fusing technology and creativity to meet the needs of diverse audiences.

I feel we need innovative ideas and new perspectives, but what do you think?

Prompt engineering style questions
You are an expert in innovative filmmaking. Please suggest ideas for developing new movie genres and storytelling techniques, considering the following points:

1. Objective: To create new movie experiences that go beyond traditional film concepts and meet the needs of diverse audiences.
2. Elements to consider:
   (a) Utilization of latest technologies (AI, VR, AR, etc.)
   (b) Introduction of interactivity
   (c) Universal approach for global markets
   (d) Direct connection with audience emotions and thoughts

3. Constraints:
   (a) Technical feasibility
   (b) Ethical and legal considerations
   (c) Coexistence with traditional movie experiences
4. Expected outcomes:
   (a) Proposal of at least 3 innovative movie genres or storytelling techniques
   (b) Specific implementation methods for each proposal
   (c) Anticipated challenges and solutions
5. Additional instructions:
   (a) Present proposals clearly in bullet points.
   (b) Add a concise explanation of about 100 characters for each proposal.
   (c) Describe technical details and how they will change the audience experience specifically.

Based on this task, please propose innovative and feasible new movie genres or storytelling techniques.

A.2 Calculation Methods

Detailed definitions, calculation methods and implementation of the evaluation indicators used in this study are described. Each indicator is automatically calculated from the textual data of the ideas using natural language processing techniques.

A.2.1 creativity

The creativity score aims to assess the originality, innovation and complexity of an idea. It is calculated from the following elements

1. Vocabulary diversity: the ratio of the number of unique words used in an idea to the total number of words. The use of a diverse vocabulary suggests creative thinking.
2. use of innovative terms: frequency of occurrence of terms such as ‘innovative’, ‘breakthrough’ and ‘novel’. These terms indicate the innovative nature of ideas.
3. Complex sentence structure: the proportion of sentences containing subordinate clauses or multiple clauses. Complex sentence structure reflects the complexity and creativity of an idea.

```
def creativity_score(text):
    doc = nlp(text)
    sentences = list(doc.sents)
    words = [token.text for token in doc if not token.is_stop
        and token.is_alpha]
    unique_words = len(set(words))
    total_words = len(words)
    vocab_diversity = unique_words / total_words if
       total_words != 0 else 0
```

```
    specialized_terms = ['innovative', 'breakthrough', '
        revolutionary', 'novel', 'unique', 'creative', '
        original']
    specialized_count = sum([1 for word in words if word.
        lower() in specialized_terms])

    complex_sentences = sum([1 for sent in sentences if len(
        list(sent.root.children)) > 3])
    sentence_complexity = complex_sentences / len(sentences)
        if len(sentences) != 0 else 0

    return vocab_diversity + (specialized_count / total_words
         if total_words != 0 else 0) + sentence_complexity
```

A.2.2 practicality

The utility score measures the viability, usefulness and specific action orientation of an idea. It is calculated from the following elements

1. Practical keywords: the frequency of occurrence of keywords such as ‘implement’, ‘feasible’ and ‘useful’. These keywords suggest the practicality of an idea.
2. Practical phrases: frequency of occurrence of practical phrases such as ‘cost-effective solution’, ‘practical approach’, etc.
3. Verb variety: the number of unique verbs in the idea. Verb variety indicates orientation towards specific actions.

```
practical_keywords = [
    'implement', 'feasible', 'useful', 'effective', '
        efficient',
    'scalable', 'deploy', 'execute', 'measure', 'resource',
    'constraint', 'budget', 'cost', 'timeframe', 'deadline',
    'risk', 'benefit', 'roi', 'advantage', 'disadvantage',
    'impact', 'outcome', 'result', 'solution', 'problem',
    'challenge', 'opportunity', 'alternative', 'trade-off',
    'practical', 'applicable', 'workable', 'viable', '
        operational',
    'functional', 'pragmatic', 'realistic', 'achievable', '
        doable']

def practical_score(text):
    doc = nlp(text)
    tokens = [token.lemma_.lower() for token in doc if not
        token.is_stop and token.is_alpha]
    practical_count = sum([1 for token in tokens if token in
        practical_keywords])
```

```
    practical_phrases = sum([1 for i in range(len(tokens)-1)
        if tokens[i] in practical_keywords and tokens[i+1] in
        practical_keywords])

    verbs = [token.lemma_ for token in doc if token.pos_ == "
        VERB"]
    action_verbs = len(set(verbs))

    return (practical_count + practical_phrases +
        action_verbs) / len(tokens) if len(tokens) != 0 else 0
```

A.2.3 concreteness

The concreteness score assesses the level of detail, clarity and real-world relevance of an idea. It is calculated from the following elements

1. Concrete keywords: frequency of occurrence of keywords such as ‘detailed’, ‘specific’ and ‘concrete’. These keywords suggest the specificity of the idea.
2. Concrete paragraphs: the percentage of paragraphs containing specific keywords.
3. Numerical values and specific expressions: frequency of occurrence of numerical values and specific expressions (e.g. organisation name, place name). These expressions indicate the concreteness of the idea and its relevance to the real world.

```
specificity_keywords = [
    'detailed', 'specific', 'precise', 'explicit', 'exact',
    'concrete', 'particular', 'definite', 'clear-cut', '
        unambiguous',
    'quantitative', 'measurable', 'data', 'statistics', '
        figures',
    'numbers', 'metrics', 'kpi', 'benchmark', 'criterion',
    'parameter', 'indicator', 'specification', 'timeline', '
        schedule',
    'milestone', 'phase', 'step', 'procedure', 'protocol',
    'methodology', 'technique', 'approach', 'framework', '
        structure',
    'outline', 'blueprint', 'roadmap', 'plan', 'strategy'
]

def specificity_score(text):
    doc = nlp(text)
    tokens = [token.lemma_.lower() for token in doc if not
        token.is_stop and token.is_alpha]
    specificity_count = sum([1 for token in tokens if token
        in specificity_keywords])

    paragraphs = text.split('\n')
```

```
    specific_paragraphs = sum([1 for paragraph in paragraphs
        if any(keyword in paragraph.lower() for keyword in
        specificity_keywords)])

    numbers = len([token for token in doc if token.like_num])
    named_entities = len([ent for ent in doc.ents if ent.
        label_ in ['PERSON', 'ORG', 'GPE', 'PRODUCT']])

    total_words = len(tokens)
    return (specificity_count + specific_paragraphs + numbers
         + named_entities) / (total_words + len(paragraphs))
        if total_words != 0 else 0
```

A.2.4 interactive naturalness

The interactive naturalness score measures how natural an idea feels in an interactive context. It is calculated from the following elements

1. Dialogic expressions: the frequency of the occurrence of dialogic expressions such as 'you', 'your' and 'let's'. These expressions indicate the dialogical nature of the idea.
2. diversity of sentence types: the proportion of different types of sentences, such as platitudes, interrogatives, exclamations, etc. The variety of sentence types reflects the natural flow of dialogue.
3. Variety of emotions: the variety of emotions expressed throughout the text. Variation in emotions is observed in natural dialogue.
4. Variation in sentence length: standard deviation of sentence length. In natural dialogue, there is variation in sentence length.

```
def dialogue_naturalness_score(text):
    doc = nlp(text)

    # List of dialogic expressions
    dialogue_expressions = ['you', 'your', 'we', 'our', 'let
        \'s', 'shall we', 'what do you think', 'how about']

    # Variety of sentence types (e.g. platitudes,
        interrogatives, exclamations)
    sentence_types = set([sent.root.tag_ for sent in doc.
        sents])
    sentence_diversity = len(sentence_types) / len(list(doc.
        sents))

    # Frequency of use of dialogic expressions
    dialogue_expr_count = sum([1 for token in doc if token.
        text.lower() in dialogue_expressions])
```

```
    # Variation in sentence lengths (sentence lengths tend to
        vary in natural dialogues)
    sentence_lengths = [len(sent) for sent in doc.sents]
    length_variance = np.var(sentence_lengths) if len(
        sentence_lengths) > 1 else 0

    # Variance of sentiments (using TextBlob)
    blob = TextBlob(text)
    sentiment_scores = [sentence.sentiment.polarity for
        sentence in blob.sentences]
    sentiment_diversity = np.std(sentiment_scores) if len(
        sentiment_scores) > 1 else 0

    # Calculate the score (weight each element and sum it)
    score = (
        0.3 * sentence_diversity +
        0.3 * (dialogue_expr_count / len(doc)) +
        0.2 * min(1, length_variance / 100) +
        0.2 * sentiment_diversity
    )

    return score
```

A.2.5 facilitating thinking

The Facilitating Thinking score assesses the potential of an idea to facilitate critical and analytical thinking. It is calculated from the following elements

1. Open-ended questions: the percentage of questions that begin with ‘what’, ‘why’ or ‘how’. Open-ended questions promote deeper thinking.
2. expressions that promote thinking: frequency of occurrences of expressions such as ‘consider’, ‘analyse’ and ‘reflect’. These expressions promote critical thinking.
3. Complex sentence structures: the proportion of sentences containing subordinate or modifying clauses. Complex sentence structures reflect advanced thinking.
4. Words representing abstract thinking: frequency of occurrence of words such as ‘concept’, ‘theory’ and ‘perspective’. These words indicate abstract thinking.

```
def thought_promotion_score(text):
    doc = nlp(text)

    # List of thought-promoting expressions
    thought_promoting_expressions = ['why', 'how', 'what if',
        'consider', 'imagine', 'think about', 'reflect on', '
        analyze']

    # Number of open-ended questions
```

```
    open_ended_questions = sum([1 for sent in doc.sents if
        sent.text.strip().endswith('?') and not sent.text.
        lower().startswith(('is', 'are', 'do', 'does', 'has',
        'have', 'can', 'could', 'will', 'would'))])

    # Frequency of use of thought-promoting expressions
    thought_expr_count = sum([1 for token in doc if any(expr
        in token.text.lower() for expr in
        thought_promoting_expressions)])

    # Percentage of complex sentence structures (percentage
        of sentences containing dependent clauses)
    complex_sentences = sum([1 for sent in doc.sents if any(
        token.dep_ in ['advcl', 'acl', 'relcl'] for token in
        sent)])
    complex_sentence_ratio = complex_sentences / len(list(doc
        .sents))

    # Use words for abstract concepts and higher-order
        thinking
    abstract_thinking_words = ['concept', 'theory', '
        hypothesis', 'analysis', 'synthesis', 'evaluation', '
        perspective', 'implication']
    abstract_word_count = sum([1 for token in doc if token.
        lemma_.lower() in abstract_thinking_words])

    # Calculate score (weight each element and sum)
    score = (
         0.3 * (open_ended_questions / len(list(doc.sents))) +
         0.3 * (thought_expr_count / len(doc)) +
         0.2 * complex_sentence_ratio +
         0.2 * (abstract_word_count / len(doc))
    )

    return score
```

A.2.6 complexity

The complexity score assesses the structural complexity of an idea. It is calculated from the following factors

1. Sentence depth: the average depth of subordinate clauses and nested clauses. A deeper structure indicates the complexity of the idea.
2. Dependent clauses: the proportion of dependent clauses. The number of subordinate clauses reflects the complexity of the idea.

```
def complexity_score(doc):
    sentence_depths = [len(list(sent.root.ancestors)) for
        sent in doc.sents]
    avg_depth = sum(sentence_depths) / len(sentence_depths)
        if sentence_depths else 0
    subordinate_clauses = len([token for token in doc if
        token.dep_ == "advcl"])
    return (avg_depth + subordinate_clauses) / len(doc)
```

A.2.7 technicality

The technicality score measures the extent to which an idea contains technical or professional content. It is calculated from the following elements

1. Technical specific expressions: frequency of occurrence of specific expressions such as ‘ORG’, ‘PRODUCT’, ‘GPE’, etc. These eigenexpressions suggest technical or specialised content.

```
def technicality_score(doc):
    entities = [ent.label_ for ent in doc.ents]
    technical_entities = [e for e in entities if e in ['ORG',
         'PRODUCT', 'GPE', 'LAW', 'WORK_OF_ART']]
    return len(technical_entities) / len(doc)
```

A.2.8 diversity

The diversity score assesses the diversity and richness of the vocabulary of ideas. It is calculated from the following factors:

1. Vocabulary diversity: the variance of the frequency of use of words in the text. High variance indicates lexical diversity.
2. Type/token ratio: the ratio between the number of unique words (types) and the total number of words (tokens). A high ratio indicates lexical richness.

```
def diversity_score(doc):
    word_freq = Counter([token.text.lower() for token in doc
        if not token.is_stop and token.is_alpha])
    total_words = sum(word_freq.values())
    word_entropy = -sum((count / total_words) * math.log2(
        count / total_words) for count in word_freq.values())
    return word_entropy / math.log2(len(word_freq)) if
        word_freq else 0
```

A.2.9 consistency

The consistency score assesses the overall coherence and logical flow of ideas. It is calculated from the following factors

1. Lexical overlap of adjacent sentences: the proportion of words shared between adjacent sentence pairs. A high percentage indicates coherence.

```
def coherence_score(doc):
    sentences = list(doc.sents)
    coherence = sum(len(set(sent1.lemma_) & set(sent2.lemma_)
        )
                    for sent1, sent2 in zip(sentences,
                        sentences[1:]))
    return coherence / (len(sentences) - 1) if len(sentences)
         > 1 else 0
```

A.2.10 Readability

The readability score evaluates the degree to which an idea is easy to read and understand. It is calculated from the following factors:

1. Sentence length: the average number of words in a sentence. Shorter sentences improve readability.
2. Number of syllables: the average number of syllables in a word. Shorter words improve readability.

```
def readability_score(doc):
    words = [token.text for token in doc if not token.
        is_punct]
    sentences = list(doc.sents)
    avg_sentence_length = len(words) / len(sentences)
    avg_syllables_per_word = sum(len([char for char in word
        if char.lower() in 'aeiou']) for word in words) / len(
        words)
    return 206.835 - 1.015 * avg_sentence_length - 84.6 *
        avg_syllables_per_word
```

A.2.11 Density of proper nouns

Density of proper nouns measures the number of proper nouns (such as names of people, organisations and places) within an idea.

```
len(doc.ents) / len(doc)
```

A.2.12 Density of parts of speech

Density of parts of speech measures the number of nouns, verbs, adjectives and adverbs within an idea. These parts of speech provide insight into the content and style of an idea.

```
len([token for token in doc if token.pos_ in ['NOUN', 'VERB',
     'ADJ', 'ADV']]) / len(doc)
```

A.2.13 Average word length

Average word length measures the average number of characters in a word in an idea.

```
sum(len(token.text) for token in doc if not token.is_punct) /
    len([token for token in doc if not token.is_punct]),
```

A.2.14 Lexical diversity (type/token ratio)

Lexical diversity measures the ratio of the number of unique words (types) to the total number of words (tokens). This index indicates the diversity and richness of the text's vocabulary.

```
len(set([token.text.lower() for token in doc if not token.
    is_punct])) / len([token for token in doc if not token.
    is_punct])
```

A.2.15 Dependency Distance

Dependency distance measures the average distance between a word and its syntactic head. This indicates the complexity of the syntactic structure of an idea and the strength of the relationship between words.

```
sum(abs(token.i - token.head.i) for token in doc if token.
    dep_ ! = 'ROOT') / len([token for token in doc if token.
    dep_ != 'ROOT'])
```

A.2.16 Frequency of Passive Voice Usage

The frequency of passive voice usage measures the proportion of sentences that use the passive voice. The passive voice suggests that the focus of the idea is on the action itself, rather than the subject of the action.

```
len([1 for token in doc if token.dep_ == 'nsubjpass']) / len(
    list(doc.sents))
```

A.3 List of the top 20 frequently occurring words

ans 1	Coral word	counts	ChatGPT4 word	counts	Gemini 1.0 pro word	counts	Gemini 1.5 flash word	counts	Gemini 1.5 pro word	counts	Claude Haiku 3 word	counts	Claude Sonnet 3.5 word	counts	Claude opus 3 word	counts
1	student	356	student	410	student	302	learning	443	learning	347	education	179	learning	221	learning	274
2	learning	333	learning	315	educational	285	need	434	student	294	need	168	need	196	need	254
3	education	246	education	253	education	218	student	356	need	263	change	158	education	194	student	217
4	change	235	technology	194	learning	212	education	303	education	204	educational	150	skill	182	education	203
5	skill	211	need	171	challenge	205	technology	246	world	195	student	142	student	163	skill	190
6	social	197	educational	169	change	193	skill	211	skill	177	challenge	127	like	163	change	149
7	educational	182	school	167	value	162	change	201	technology	175	system	125	change	141	educational	135
8	provide	178	teacher	166	skill	139	critical	179	change	164	learning	102	technology	123	important	129
9	thinking	166	mental	165	system	131	world	158	future	162	complex	93	human	117	social	125
10	value	161	role	163	need	128	thinking	157	critical	156	critical	91	challenge	109	challenge	121
11	teacher	152	health	157	technology	127	collaboration	157	thinking	141	skill	91	model	109	like	118
12	technology	149	physical	153	core	123	future	155	global	136	like	90	role	105	technology	116
13	advancement	146	change	153	technological	122	value	145	teacher	125	issue	89	value	91	learn	108
14	global	145	learn	142	mental	122	system	140	collaboration	121	model	88	assessment	91	role	105
15	support	144	value	138	health	120	challenge	136	challenge	120	value	88	teacher	89	emotional	102
16	role	143	support	133	teacher	114	physical	134	foster	113	teacher	88	thinking	86	model	100
17	critical	139	skill	129	thinking	112	mental	125	mental	109	traditional	86	future	83	future	97
18	mental	139	provide	129	advancement	111	creativity	122	value	106	social	84	system	81	teacher	96
19	technological	138	system	124	support	111	health	122	health	105	technology	84	critical	80	evolve	89
20	health	137	traditional	123	global	107	role	121	problem	102	role	84	physical	79	believe	86

ans 2	Coral word	counts	ChatGPT4 word	counts	Gemini 1.0 pro word	counts	Gemini 1.5 flash word	counts	Gemini 1.5 pro word	counts	Claude Haiku 3 word	counts	Claude Sonnet 3.5 word	counts	Claude opus 3 word	counts
1	student	1270	learning	565	learning	786	learning	1732	learning	1341	learning	1039	learning	1435	learning	950
2	learning	1072	technology	513	technology	600	technology	891	technology	727	student	856	technology	858	student	635
3	technology	896	global	450	student	600	student	880	student	582	technology	851	skill	562	challenge	585
4	skill	683	education	446	educational	415	education	650	education	495	educational	572	educational	560	technology	576
5	opportunity	621	educational	405	education	409	opportunity	568	skill	413	opportunity	561	opportunity	516	development	571
6	global	543	development	363	global	401	skill	568	global	413	base	506	education	507	education	555
7	educational	477	student	327	potential	372	global	545	challenge	398	education	505	global	505	opportunity	536
8	potential	470	future	315	development	369	challenge	514	future	393	challenge	500	student	490	skill	490
9	school	454	recommendation	300	current	334	personalized	449	opportunity	386	skill	470	potential	475	potential	405
10	education	451	current	297	opportunity	322	thinking	419	personalized	367	potential	444	challenge	446	global	400
11	challenge	439	mental	286	skill	293	potential	411	development	361	school	443	base	444	current	371
12	thinking	423	health	284	recommendation	280	digital	387	world	331	need	443	development	412	school	345
13	critical	401	skill	261	mental	273	need	382	potential	330	development	414	current	385	recommendation	344
14	change	396	model	213	support	267	critical	375	current	324	global	411	digital	371	provide	342
15	support	375	digital	208	health	267	current	350	mental	317	year	386	thinking	347	educational	314
16	health	372	need	203	value	239	mental	338	health	309	teacher	366	recommendation	345	educator	301
17	mental	371	teacher	203	future	237	health	330	need	308	digital	353	job	333	base	301
18	need	368	ai	201	need	231	provide	328	recommendation	304	current	333	citizenship	326	c	297
19	current	367	situation	193	model	231	future	326	critical	298	job	326	need	315	b	296
20	social	346	curriculum	183	challenges	227	ai	306	experience	290	support	312	personalized	315	support	294

Table 14 List of the top 20 frequently occurring words extracted from eight AI models on the theme 'Future of Education'.

Table 15 *

Color coding in the table: Blue: Words common to both Answer 1 and Answer 2, with similar frequency rankings Green: Words present in both answers but with differing frequency rankings Orange: Words predominantly found in Answer 1 Purple: Words predominantly found in Answer 2

ans 1	Coral word	counts	ChatGPT4 word	counts	Gemini 1.0 pro word	counts	Gemini 1.5 flash word	counts	Gemini 1.5 pro word	counts	Claude Haiku 3 word	counts	Claude Sonnet 3.5 word	counts	Claude opus 3 word	counts
1	employee	495	communication	557	communication	565	communication	590	communication	415	communication	280	communication	289	communication	346
2	communication	490	employee	319	employee	363	department	215	project	191	employee	263	employee	237	employee	285
3	encourage	249	department	305	encourage	226	employee	206	employee	177	encourage	210	collaboration	219	collaboration	249
4	team	232	team	266	foster	205	share	199	information	172	collaboration	173	department	183	encourage	223
5	help	228	help	214	create	190	information	197	share	162	information	156	cross	157	department	192
6	company	224	tool	189	information	174	team	186	feedback	162	team	156	organization	155	organism	189
7	department	221	project	170	share	170	foster	181	department	154	organization	154	team	150	create	172
8	collaboration	214	different	158	team	134	encourage	178	foster	150	share	146	different	149	foster	170
9	foster	196	company	152	collaboration	128	project	176	encourage	146	department	143	encourage	144	company	157
10	work	187	feedback	151	different	126	open	161	work	143	foster	140	information	139	share	155
11	create	177	work	146	department	120	create	155	organism	139	cross	128	foster	135	live	142
12	feedback	152	regular	133	project	101	collaboration	153	create	137	help	127	functional	129	information	142
13	culture	149	cross	126	generational	101	different	135	team	123	knowledge	118	knowledge	126	promote	129
14	project	148	collaboration	120	sharing	100	feedback	131	development	120	sharing	112	promote	126	organization	129
15	information	148	implement	118	knowledge	96	culture	119	knowledge	120	implement	109	implement	121	culture	120
16	provide	138	strategy	116	facilitate	95	generational	110	different	119	work	107	sharing	118	work	112
17	organization	129	encourage	111	feedback	93	need	107	generational	117	culture	106	generational	105	idea	109
18	share	128	like	107	culture	91	knowledge	104	company	113	approach	106	organism	103	help	109
19	implement	125	generational	103	promote	90	understanding	102	cross	112	facilitate	94	share	102	cross	106
20	approach	122	challenge	102	provide	89	work	102	sale	112	project	87	help	99	knowledge	104

ans 2	Coral word	counts	ChatGPT4 word	counts	Gemini 1.0 pro word	counts	Gemini 1.5 flash word	counts	Gemini 1.5 pro word	counts	Claude Haiku 3 word	counts	Claude Sonnet 3.5 word	counts	Claude opus 3 word	counts
1	communication	417	communication	400	communication	559	communication	697	communication	508	communication	574	communication	548	communication	414
2	information	313	challenge	309	information	234	challenge	297	implementation	262	information	195	promote	190	collaboration	188
3	employee	299	implementation	295	employee	229	project	264	challenge	255	employee	194	implement	176	employee	143
4	proposal	295	effect	253	expect	167	implementation	258	expect	250	implement	192	employee	167	term	141
5	team	265	term	244	effect	164	effect	249	effect	250	project	173	knowledge	162	project	139
6	implementation	260	project	184	implement	146	information	245	project	245	cross	160	department	158	cross	130
7	effect	252	explanation	173	collaboration	146	expect	234	explanation	216	sharing	159	information	149	implement	128
8	provide	233	expect	157	implementation	144	team	231	department	214	encourage	156	collaboration	140	share	116
9	ensure	233	department	145	explanation	140	countermeasure	229	countermeasure	210	platform	138	cross	134	knowledge	116
10	encourage	230	countermeasure	132	sharing	135	knowledge	224	knowledge	176	foster	133	sharing	134	information	113
11	challenge	229	short	131	foster	135	explanation	220	character	165	knowledge	129	training	132	team	110
12	department	188	employee	120	provide	127	improve	215	improve	154	culture	126	team	127	internal	102
13	knowledge	179	proposal	117	knowledge	124	program	210	information	141	collaboration	122	project	126	functional	100
14	enhance	176	ensure	116	establish	124	platform	196	platform	138	functional	110	foster	120	provide	98
15	project	174	timeframe	112	cross	118	training	195	employee	130	department	110	encourage	108	improve	97
16	sharing	174	feedback	112	platform	117	term	182	sharing	129	establish	106	term	106	company	96
17	improve	166	tool	105	create	105	sharing	173	term	127	facilitate	104	open	96	sharing	94
18	platform	161	platform	102	reduce	104	department	156	training	121	promote	103	generational	95	info	85
19	company	156	role	101	project	103	clear	156	program	120	management	102	understanding	93	strategy	85
20	countermeasure	148	liaison	99	team	98	ensure	149	provide	118	program	97	social	92	department	82

Table 16 List of the top 20 frequently occurring words extracted from eight AI models on the theme 'Improving Internal Corporate Communication'.

Table 17 *

Color coding in the table: Blue: Words common to both Answer 1 and Answer 2, with similar frequency rankings Green: Words present in both answers but with differing frequency rankings Orange: Words predominantly found in Answer 1 Purple: Words predominantly found in Answer 2

ans 1	Coral		ChatGPT4		Gemini 1.0 pro		Gemini 1.5 flash		Gemini 1.5 pro		Claude Haiku 3		Claude Sonnet 3.5		Claude opus 3	
	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts
1	meal	583	meal	470	meal	463	meal	374	meal	340	meal	415	meal	360	meal	338
2	time	273	cooking	443	cooking	302	cooking	265	cooking	268	wife	217	family	173	healthy	201
3	cooking	267	healthy	225	family	232	healthy	203	time	238	cooking	171	cooking	155	wife	190
4	family	256	cook	194	time	205	recipe	180	healthy	197	family	170	healthy	121	time	186
5	wife	218	help	186	wife	178	time	179	family	196	healthy	144	time	112	cooking	177
6	help	182	time	175	cook	150	wife	170	recipe	176	simple	115	work	96	help	172
7	eating	178	like	171	healthy	147	family	161	wife	172	help	106	simple	93	family	169
8	habit	176	simple	170	health	118	cook	158	cook	171	like	96	week	89	work	138
9	healthy	175	family	167	preparation	99	work	136	solution	127	cook	91	like	87	prepare	119
10	health	166	wife	154	consider	97	small	134	start	120	pot	90	help	83	effort	101
11	improve	157	kitchen	139	simple	93	focus	130	small	120	involve	89	involve	82	recipe	99
12	simple	157	skill	122	prepare	91	help	123	work	115	work	87	wife	80	simple	95
13	preparation	141	consider	120	ask	89	solution	115	health	113	week	85	habit	78	cooker	93
14	recipe	136	preparation	107	task	88	start	113	change	112	prepare	79	cook	76	vegetable	91
15	task	131	health	102	help	86	simple	112	weekend	105	slow	79	weekend	70	change	88
16	prepare	128	prepare	101	support	83	find	106	focus	98	cooker	79	responsibility	68	week	86
17	cook	123	clean	98	challenge	77	feel	101	simple	96	time	79	change	68	like	85
18	consider	122	week	96	planning	76	ingredient	101	effort	93	nutritious	77	prep	67	habit	85
19	nutritious	113	cooker	91	block	75	week	100	week	91	instant	76	service	64	health	84
20	skill	110	ingredient	84	recipe	74	like	96	ingredient	91	skill	75	eating	64	consider	83

ans 2	Coral		ChatGPT4		Gemini 1.0 pro		Gemini 1.5 flash		Gemini 1.5 pro		Claude Haiku 3		Claude Sonnet 3.5		Claude opus 3	
	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts
1	meal	685	meal	550	meal	405	meal	840	meal	486	meal	549	meal	679	meal	664
2	cooking	349	cooking	516	cooking	322	family	448	cook	306	family	289	cooking	396	family	337
3	veggie	348	cook	320	cook	290	cooking	447	cooking	304	cooking	272	family	371	cooking	288
4	family	304	family	318	family	238	vegetable	375	family	277	vegetable	251	recipe	190	time	279
5	cook	253	vegetable	307	vegetable	197	recipe	350	vegetable	268	time	184	time	181	cook	225
6	time	239	time	281	cooker	166	time	341	recipe	229	prepare	176	skill	144	vegetable	196
7	skill	221	skill	243	slow	163	simple	256	chop	226	easy	150	preparation	140	recipe	192
8	simple	192	cooker	184	recipe	150	ingredient	224	time	191	pot	140	cook	136	ingredient	185
9	quick	159	recipe	173	sauce	134	cook	209	veggie	160	busy	127	prepare	134	save	150
10	healthy	155	preparation	168	chicken	121	easy	202	chicken	155	nutritious	122	vegetable	129	strategy	148
11	vegetable	154	simple	153	time	120	prepare	200	pasta	135	recipe	116	like	128	grain	140
12	chop	153	slow	152	member	118	healthy	188	easy	132	chicken	113	protein	121	skill	138
13	pasta	151	like	146	pasta	112	like	184	week	130	simple	113	involve	114	prepare	128
14	recipe	148	chicken	133	skill	112	busy	165	healthy	130	slow	108	gradually	111	easy	121
15	pot	147	use	133	pan	105	preparation	162	pan	119	cooker	108	ingredient	110	simple	121
16	easy	146	explanation	130	class	104	skill	160	small	117	strategy	107	task	110	preparation	120
17	slow	146	ingredient	125	simple	102	save	155	celebrate	112	protein	106	cooker	109	dinner	100
18	save	143	week	124	task	100	chicken	146	start	108	grain	94	slow	107	nutritious	100
19	improve	142	pre	124	weekend	97	nutritious	145	simple	108	skill	89	portion	104	protein	99
20	like	141	prepare	117	ingredient	95	quick	145	quick	108	roasted	86	member	101	busy	96

Table 18 List of the top 20 frequently occurring words extracted from eight AI models on the theme 'Healthy family meal planning'.

Table 19 *

Color coding in the table: Blue: Words common to both Answer 1 and Answer 2, with similar frequency rankings Green: Words present in both answers but with differing frequency rankings Orange: Words predominantly found in Answer 1 Purple: Words predominantly found in Answer 2

ans 1	Coral		ChatGPT4		Gemini 1.0 pro		Gemini 1.5 flash		Gemini 1.5 pro		Claude Haiku 3		Claude Sonnet 3.5		Claude opus 3	
	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts
1	audience	822	film	370	audience	428	audience	447	film	378	audience	443	audience	356	audience	385
2	storytelling	494	viewer	345	storytelling	336	experience	359	audience	367	storytelling	273	storytelling	312	storytelling	267
3	experience	428	storytelling	302	experience	230	film	351	storytelling	256	experience	270	experience	286	story	255
4	create	374	audience	300	create	228	storytelling	325	experience	235	narrative	215	narrative	234	experience	243
5	movie	360	story	279	movie	171	create	266	ai	173	new	190	create	168	movie	215
6	story	317	experience	228	interactive	154	ai	193	narrative	169	explore	159	story	150	film	187
7	diverse	297	narrative	199	technology	147	story	187	story	163	create	145	technology	132	new	156
8	film	240	create	168	narrative	130	interactive	180	create	148	technology	138	emotion	120	viewer	156
9	technology	224	ai	163	immersive	130	narrative	176	explore	146	diverse	114	movie	120	narrative	142
10	narrative	219	movie	156	technique	125	new	174	technology	140	element	102	ai	114	create	135
11	interactive	215	new	155	ai	120	technology	148	interactive	130	cultural	102	cultural	114	ai	130
12	develop	207	element	149	explore	117	explore	147	viewer	124	technique	98	new	105	technology	128
13	immersive	180	base	141	cultural	115	embrace	132	imagine	119	approach	98	explore	104	idea	106
14	content	169	content	140	diverse	105	vr	128	vr	119	story	96	idea	103	interactive	104
15	explore	167	character	139	emotional	95	viewer	124	visual	114	interactive	92	interactive	102	diverse	99
16	cultural	165	global	137	film	94	character	122	world	104	challenge	91	develop	92	character	99
17	idea	164	interactive	137	perspective	92	diverse	118	new	99	develop	88	thought	90	emotion	99
18	technique	158	real	137	story	89	immersive	113	character	99	idea	87	emotional	89	emotional	94
19	involve	154	develop	135	vr	87	visual	113	cultural	94	thought	85	immersive	87	medium	89
20	viewer	151	cultural	125	viewer	87	allow	99	challenge	91	visual	83	incorporate	74	explore	86

ans 2	Coral		ChatGPT4		Gemini 1.0 pro		Gemini 1.5 flash		Gemini 1.5 pro		Claude Haiku 3		Claude Sonnet 3.5		Claude opus 3	
	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts	word	counts
1	audience	413	viewer	289	audience	183	audience	406	experience	327	audience	469	audience	293	viewer	283
2	experience	292	SFF	278	experience	166	experience	398	audience	307	experience	405	narrative	244	story	219
3	movie	266	challenge	262	viewer	148	narrative	329	narrative	266	movie	286	ai	238	ai	198
4	narrative	253	film	249	proposal	137	ai	300	film	224	narrative	271	storytelling	165	movie	167
5	proposal	246	experience	194	narrative	137	film	201	technical	206	cinematic	211	experience	162	ar	146
6	create	224	emotional	192	implementation	135	technical	193	viewer	176	storytelling	191	vr	120	challenge	143
7	interactive	171	proposal	184	create	111	viewer	172	ai	167	technology	184	emotion	113	experience	142
8	emotion	159	audience	184	immersive	98	vr	171	real	157	create	178	emotional	113	film	130
9	global	156	narrative	158	interactive	85	storytelling	170	ar	152	viewer	178	ar	108	audience	127
10	implementation	154	ar	154	vr	84	challenge	166	vr	129	ai	153	immersive	106	vr	117
11	technical	150	real	152	ai	84	solution	161	solution	129	innovative	147	interactive	97	storytelling	111
12	viewer	147	implementation	151	cinema	83	real	160	challenge	118	implementation	147	base	92	real	100
13	emotional	144	ai	151	film	83	implementation	159	emotional	108	traditional	140	real	90	interactive	98
14	story	137	datum	148	technical	82	story	151	character	99	emotional	133	create	88	create	85
15	cinema	135	vr	144	storytelling	81	create	150	storytelling	98	ensure	117	storyline	87	innovative	77
16	immersive	132	time	140	challenge	76	ar	149	interactive	96	technique	116	datum	83	generate	73
17	ai	132	interactive	136	character	73	technology	143	time	95	real	110	global	80	time	70
18	cultural	127	technology	129	genre	72	time	135	element	95	challenge	109	time	78	narrative	69
19	diverse	125	technical	122	story	68	emotional	129	details	88	story	109	biometric	74	personalized	67
20	storytelling	119	method	121	emotional	67	explanation	127	datum	86	time	105	utilize	73	new	65

Table 20 List of the top 20 frequently occurring words extracted from eight AI models on the theme 'New films'.

Table 21 *

Color coding in the table: Blue: Words common to both Answer 1 and Answer 2, with similar frequency rankings Green: Words present in both answers but with differing frequency rankings Orange: Words predominantly found in Answer 1 Purple: Words predominantly found in Answer 2

Table 22 Metrics for Various Models, The Future of Education 1

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.71	0.07	0.41	0.06	0.73	0.09	0.52	0.10	0.62	0.13	0.35	0.13	0.46	0.09	0.19	0.07
practicality	0.55	0.08	0.23	0.04	0.53	0.08	0.29	0.07	0.61	0.11	0.24	0.06	0.54	0.11	0.20	0.06
specificity	0.09	0.04	0.26	0.05	0.23	0.05	0.28	0.06	0.10	0.06	0.18	0.06	0.13	0.06	0.27	0.05
dialogue	0.52	0.15	0.54	0.05	0.52	0.14	0.61	0.05	0.65	0.15	0.51	0.16	0.56	0.10	0.44	0.06
thought	0.55	0.13	0.65	0.11	0.51	0.11	0.41	0.11	0.33	0.15	0.28	0.16	0.24	0.11	0.18	0.08
complexity	0.45	0.11	0.56	0.11	0.43	0.13	0.38	0.11	0.41	0.14	0.25	0.11	0.25	0.12	0.22	0.09
technicality	0.13	0.06	0.29	0.08	0.19	0.07	0.26	0.08	0.13	0.09	0.12	0.09	0.17	0.09	0.19	0.07
diversity	0.74	0.05	0.28	0.05	0.72	0.04	0.44	0.05	0.71	0.06	0.33	0.07	0.64	0.08	0.17	0.06
coherence	0.71	0.06	0.83	0.09	0.61	0.07	0.47	0.09	0.30	0.16	0.38	0.20	0.29	0.14	0.20	0.09
readability	0.46	0.08	0.35	0.12	0.55	0.05	0.52	0.08	0.56	0.08	0.50	0.12	0.76	0.09	0.58	0.06
named	0.05	0.03	0.23	0.04	0.20	0.03	0.31	0.12	0.08	0.05	0.14	0.07	0.10	0.05	0.19	0.05
lexical	0.57	0.05	0.47	0.04	0.52	0.06	0.35	0.08	0.30	0.11	0.35	0.06	0.21	0.08	0.25	0.07
avg. word	0.61	0.09	0.62	0.09	0.50	0.08	0.57	0.10	0.62	0.12	0.69	0.09	0.31	0.14	0.70	0.10
type token	0.60	0.07	0.06	0.03	0.63	0.05	0.36	0.06	0.63	0.08	0.21	0.07	0.49	0.08	0.09	0.04
dependency	0.23	0.07	0.38	0.10	0.22	0.04	0.29	0.20	0.11	0.05	0.26	0.21	0.09	0.05	0.17	0.05
passive	0.22	0.16	0.24	0.10	0.14	0.09	0.15	0.09	0.07	0.07	0.10	0.09	0.07	0.06	0.06	0.05

A.4 Open-source toolkit Metrics for Various Models

A.5 Metrics for Various Models

Table 23 Metrics for Various Models, The Future of Education 2

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.52	0.10	0.29	0.09	0.75	0.10	0.36	0.08	0.74	0.11	0.37	0.09	0.68	0.10	0.39	0.11
practicality	0.64	0.11	0.27	0.05	0.75	0.12	0.26	0.07	0.63	0.11	0.18	0.06	0.62	0.11	0.29	0.08
specificity	0.12	0.06	0.23	0.06	0.21	0.08	0.37	0.07	0.16	0.08	0.40	0.07	0.12	0.06	0.51	0.19
dialogue	0.62	0.11	0.51	0.04	0.40	0.25	0.54	0.05	0.56	0.20	0.53	0.05	0.43	0.19	0.54	0.06
thought	0.29	0.13	0.28	0.08	0.58	0.18	0.49	0.13	0.58	0.15	0.53	0.12	0.43	0.15	0.41	0.12
complexity	0.23	0.13	0.30	0.09	0.33	0.20	0.36	0.10	0.41	0.17	0.36	0.10	0.35	0.15	0.31	0.09
technicality	0.12	0.09	0.18	0.05	0.08	0.04	0.19	0.06	0.11	0.10	0.32	0.12	0.10	0.07	0.38	0.23
diversity	0.74	0.07	0.29	0.05	0.91	0.04	0.32	0.04	0.85	0.06	0.29	0.05	0.81	0.06	0.36	0.06
coherence	0.30	0.11	0.27	0.10	0.74	0.08	0.63	0.14	0.73	0.05	0.64	0.11	0.69	0.05	0.68	0.16
readability	0.78	0.07	0.57	0.05	0.56	0.08	0.44	0.12	0.56	0.10	0.34	0.09	0.62	0.07	0.53	0.12
named	0.09	0.06	0.18	0.05	0.09	0.07	0.31	0.08	0.10	0.09	0.35	0.09	0.07	0.05	0.49	0.21
lexical	0.24	0.07	0.30	0.07	0.62	0.06	0.51	0.06	0.66	0.10	0.47	0.06	0.64	0.06	0.38	0.15
avg. word	0.27	0.11	0.64	0.08	0.35	0.14	0.53	0.09	0.38	0.18	0.73	0.11	0.28	0.11	0.48	0.11
type token	0.61	0.07	0.23	0.05	0.80	0.08	0.09	0.05	0.81	0.09	0.10	0.04	0.72	0.07	0.19	0.09
dependency	0.11	0.05	0.17	0.04	0.16	0.07	0.35	0.11	0.14	0.06	0.35	0.09	0.11	0.05	0.30	0.08
passive	0.04	0.05	0.11	0.05	0.11	0.15	0.13	0.08	0.12	0.18	0.21	0.15	0.15	0.14	0.14	0.12

Table 24 Metrics for Various Models, Improved communication within the company 1

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.61	0.10	0.58	0.12	0.65	0.15	0.54	0.13	0.44	0.11	0.50	0.18	0.48	0.08	0.24	0.12
practicality	0.46	0.10	0.34	0.08	0.47	0.10	0.45	0.15	0.40	0.11	0.42	0.17	0.37	0.12	0.22	0.11
specificity	0.24	0.08	0.30	0.09	0.53	0.08	0.61	0.13	0.28	0.15	0.37	0.19	0.35	0.09	0.49	0.12
dialogue	0.50	0.11	0.64	0.08	0.46	0.12	0.61	0.06	0.48	0.11	0.57	0.15	0.55	0.07	0.49	0.06
thought	0.53	0.11	0.67	0.17	0.52	0.11	0.31	0.11	0.33	0.09	0.36	0.18	0.30	0.08	0.14	0.06
complexity	0.35	0.13	0.54	0.17	0.37	0.09	0.27	0.10	0.38	0.12	0.33	0.15	0.33	0.12	0.14	0.08
technicality	0.05	0.08	0.07	0.04	0.19	0.11	0.34	0.12	0.16	0.15	0.16	0.12	0.22	0.13	0.35	0.11
diversity	0.54	0.08	0.51	0.09	0.60	0.09	0.48	0.08	0.49	0.13	0.45	0.15	0.61	0.09	0.32	0.09
coherence	0.61	0.11	0.73	0.10	0.54	0.13	0.37	0.12	0.30	0.11	0.43	0.21	0.30	0.07	0.13	0.12
readability	0.59	0.09	0.36	0.11	0.68	0.06	0.53	0.07	0.55	0.09	0.38	0.11	0.75	0.08	0.60	0.08
named	0.11	0.09	0.12	0.09	0.37	0.08	0.38	0.08	0.19	0.15	0.20	0.15	0.22	0.06	0.32	0.08
lexical	0.57	0.11	0.63	0.07	0.44	0.08	0.34	0.10	0.39	0.11	0.47	0.15	0.24	0.10	0.22	0.08
avg. word	0.32	0.07	0.55	0.06	0.29	0.05	0.59	0.07	0.45	0.09	0.64	0.10	0.27	0.08	0.51	0.08
type token	0.34	0.06	0.32	0.10	0.45	0.05	0.52	0.06	0.49	0.08	0.46	0.15	0.44	0.08	0.19	0.09
dependency	0.13	0.04	0.32	0.16	0.11	0.03	0.31	0.19	0.06	0.03	0.14	0.11	0.10	0.03	0.08	0.04
passive	0.28	0.20	0.14	0.15	0.24	0.17	0.07	0.09	0.06	0.07	0.03	0.10	0.05	0.06	0.04	0.06

Table 25 Metrics for Various Models, Improved communication within the company 2

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.54	0.09	0.48	0.16	0.62	0.11	0.60	0.10	0.61	0.11	0.50	0.16	0.58	0.10	0.59	0.15
practicality	0.42	0.13	0.30	0.11	0.58	0.13	0.53	0.14	0.52	0.13	0.41	0.17	0.58	0.12	0.57	0.14
specificity	0.35	0.09	0.58	0.12	0.41	0.11	0.44	0.13	0.41	0.11	0.58	0.16	0.34	0.07	0.47	0.12
dialogue	0.56	0.08	0.40	0.16	0.63	0.14	0.70	0.10	0.45	0.17	0.59	0.12	0.26	0.17	0.55	0.19
thought	0.30	0.09	0.21	0.10	0.67	0.13	0.59	0.17	0.58	0.15	0.34	0.18	0.55	0.13	0.42	0.19
complexity	0.32	0.10	0.14	0.09	0.36	0.13	0.47	0.20	0.40	0.16	0.41	0.25	0.44	0.15	0.43	0.18
technicality	0.29	0.15	0.39	0.15	0.02	0.06	0.30	0.19	0.05	0.09	0.20	0.27	0.00	0.01	0.14	0.11
diversity	0.73	0.09	0.47	0.07	0.73	0.10	0.54	0.12	0.71	0.13	0.50	0.15	0.63	0.10	0.76	0.12
coherence	0.36	0.11	0.39	0.20	0.70	0.11	0.71	0.13	0.66	0.08	0.53	0.13	0.64	0.06	0.67	0.17
readability	0.79	0.08	0.54	0.09	0.56	0.11	0.36	0.11	0.55	0.09	0.45	0.10	0.63	0.06	0.62	0.14
named	0.24	0.07	0.37	0.10	0.22	0.05	0.40	0.12	0.26	0.08	0.61	0.18	0.31	0.05	0.38	0.16
lexical	0.27	0.08	0.27	0.13	0.57	0.07	0.54	0.11	0.76	0.07	0.68	0.14	0.70	0.06	0.76	0.09
avg. word	0.26	0.07	0.62	0.12	0.28	0.07	0.55	0.08	0.35	0.08	0.58	0.10	0.31	0.06	0.31	0.09
type token	0.55	0.08	0.41	0.11	0.51	0.09	0.46	0.12	0.55	0.14	0.53	0.15	0.49	0.11	0.71	0.13
dependency	0.11	0.03	0.09	0.08	0.19	0.06	0.24	0.07	0.13	0.04	0.23	0.12	0.12	0.04	0.12	0.06
passive	0.08	0.09	0.01	0.03	0.09	0.13	0.10	0.20	0.10	0.13	0.00	0.00	0.13	0.13	0.06	0.09

Table 26 Metrics for Various Models, Healthy family meal planning 1

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.48	0.11	0.53	0.12	0.64	0.10	0.49	0.14	0.28	0.13	0.44	0.23	0.35	0.10	0.25	0.11
practicality	0.63	0.10	0.26	0.08	0.58	0.09	0.22	0.08	0.70	0.11	0.36	0.11	0.62	0.12	0.21	0.10
specificity	0.37	0.07	0.25	0.07	0.40	0.08	0.41	0.09	0.27	0.13	0.31	0.11	0.35	0.11	0.40	0.10
dialogue	0.24	0.10	0.44	0.05	0.28	0.07	0.50	0.04	0.30	0.10	0.46	0.10	0.32	0.09	0.38	0.06
thought	0.52	0.11	0.49	0.12	0.52	0.11	0.33	0.10	0.35	0.10	0.29	0.15	0.24	0.11	0.18	0.06
complexity	0.43	0.10	0.29	0.10	0.36	0.11	0.30	0.11	0.37	0.13	0.26	0.10	0.29	0.12	0.21	0.07
technicality	0.01	0.02	0.12	0.05	0.02	0.03	0.17	0.06	0.07	0.07	0.14	0.09	0.06	0.06	0.13	0.08
diversity	0.56	0.10	0.54	0.10	0.57	0.09	0.45	0.08	0.58	0.08	0.66	0.11	0.70	0.10	0.37	0.12
coherence	0.54	0.06	0.67	0.08	0.52	0.07	0.44	0.08	0.26	0.11	0.26	0.23	0.17	0.08	0.18	0.09
readability	0.66	0.05	0.59	0.06	0.66	0.05	0.65	0.06	0.71	0.07	0.67	0.08	0.83	0.06	0.72	0.05
named	0.28	0.05	0.22	0.05	0.33	0.05	0.41	0.07	0.21	0.11	0.25	0.09	0.24	0.07	0.31	0.09
lexical	0.67	0.07	0.64	0.07	0.54	0.08	0.47	0.07	0.50	0.08	0.44	0.09	0.32	0.11	0.34	0.07
avg. word	0.39	0.09	0.50	0.09	0.45	0.08	0.55	0.07	0.49	0.11	0.58	0.09	0.34	0.13	0.53	0.10
type token	0.41	0.07	0.42	0.10	0.50	0.05	0.47	0.05	0.51	0.07	0.61	0.08	0.44	0.07	0.28	0.10
dependency	0.07	0.01	0.12	0.03	0.07	0.01	0.22	0.19	0.03	0.02	0.09	0.08	0.04	0.01	0.05	0.01
passive	0.05	0.04	0.08	0.07	0.06	0.04	0.04	0.04	0.02	0.03	0.02	0.04	0.02	0.02	0.04	0.03

Table 27 Metrics for Various Models, Healthy family meal planning 2

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.39	0.11	0.28	0.10	0.58	0.14	0.59	0.16	0.44	0.19	0.53	0.14	0.61	0.09	0.49	0.19
practicality	0.66	0.12	0.18	0.08	0.62	0.11	0.33	0.12	0.70	0.13	0.29	0.08	0.68	0.14	0.34	0.07
specificity	0.35	0.11	0.48	0.11	0.41	0.09	0.45	0.11	0.48	0.13	0.32	0.08	0.36	0.09	0.51	0.17
dialogue	0.32	0.09	0.37	0.07	0.19	0.10	0.62	0.08	0.26	0.14	0.46	0.13	0.17	0.10	0.54	0.19
thought	0.19	0.10	0.12	0.05	0.54	0.16	0.55	0.18	0.45	0.16	0.48	0.18	0.63	0.12	0.42	0.16
complexity	0.22	0.10	0.19	0.08	0.29	0.15	0.42	0.12	0.36	0.12	0.36	0.12	0.63	0.16	0.33	0.12
technicality	0.07	0.08	0.15	0.08	0.08	0.07	0.31	0.20	0.05	0.05	0.12	0.09	0.05	0.05	0.08	0.07
diversity	0.72	0.09	0.62	0.10	0.74	0.09	0.63	0.08	0.81	0.10	0.52	0.11	0.70	0.09	0.53	0.10
coherence	0.19	0.08	0.17	0.08	0.59	0.06	0.60	0.17	0.51	0.11	0.66	0.11	0.59	0.05	0.68	0.22
readability	0.76	0.05	0.70	0.09	0.68	0.06	0.50	0.11	0.69	0.07	0.47	0.11	0.63	0.05	0.49	0.21
named	0.27	0.07	0.39	0.08	0.42	0.07	0.45	0.14	0.48	0.10	0.30	0.08	0.41	0.07	0.49	0.19
lexical	0.35	0.10	0.34	0.09	0.60	0.07	0.59	0.14	0.76	0.08	0.76	0.13	0.81	0.06	0.84	0.08
avg. word	0.48	0.10	0.64	0.13	0.38	0.09	0.61	0.08	0.43	0.09	0.73	0.10	0.51	0.07	0.57	0.08
type token	0.56	0.08	0.58	0.07	0.66	0.08	0.61	0.11	0.79	0.11	0.60	0.09	0.66	0.08	0.51	0.10
dependency	0.04	0.02	0.05	0.02	0.07	0.01	0.15	0.04	0.07	0.02	0.13	0.08	0.08	0.01	0.11	0.03
passive	0.01	0.01	0.02	0.02	0.09	0.06	0.18	0.11	0.05	0.04	0.04	0.06	0.05	0.05	0.12	0.19

Table 28 Metrics for Various Models, New Movie 1

model	A				B				C				D			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.54	0.08	0.61	0.12	0.66	0.09	0.67	0.09	0.47	0.12	0.56	0.20	0.45	0.07	0.32	0.13
practicality	0.48	0.06	0.58	0.11	0.54	0.07	0.69	0.13	0.50	0.09	0.55	0.13	0.48	0.09	0.40	0.12
specificity	0.26	0.08	0.30	0.08	0.37	0.07	0.29	0.08	0.38	0.16	0.27	0.13	0.24	0.10	0.37	0.10
dialogue	0.40	0.06	0.52	0.11	0.41	0.13	0.48	0.07	0.36	0.13	0.43	0.16	0.39	0.05	0.38	0.10
thought	0.66	0.11	0.78	0.13	0.65	0.10	0.61	0.16	0.44	0.11	0.39	0.22	0.31	0.09	0.22	0.13
complexity	0.34	0.07	0.63	0.17	0.40	0.10	0.40	0.13	0.38	0.17	0.37	0.26	0.27	0.09	0.25	0.15
technicality	0.06	0.03	0.14	0.07	0.04	0.03	0.08	0.05	0.08	0.06	0.04	0.05	0.09	0.06	0.15	0.07
diversity	0.31	0.11	0.54	0.14	0.59	0.08	0.52	0.08	0.54	0.11	0.67	0.11	0.51	0.10	0.32	0.15
coherence	0.69	0.06	0.74	0.07	0.60	0.08	0.65	0.10	0.32	0.11	0.33	0.23	0.27	0.06	0.17	0.10
readability	0.64	0.05	0.53	0.10	0.78	0.04	0.70	0.05	0.69	0.08	0.68	0.10	0.88	0.05	0.75	0.06
named	0.08	0.04	0.17	0.06	0.24	0.05	0.23	0.05	0.22	0.15	0.19	0.07	0.16	0.07	0.22	0.07
lexical	0.73	0.08	0.70	0.08	0.51	0.07	0.48	0.10	0.46	0.10	0.48	0.16	0.29	0.09	0.27	0.11
avg. word	0.42	0.07	0.56	0.10	0.27	0.08	0.45	0.08	0.55	0.13	0.59	0.15	0.21	0.09	0.47	0.11
type token	0.24	0.07	0.46	0.10	0.45	0.05	0.48	0.07	0.48	0.07	0.67	0.10	0.36	0.07	0.31	0.11
dependency	0.09	0.03	0.13	0.06	0.10	0.03	0.27	0.18	0.04	0.02	0.09	0.09	0.05	0.02	0.05	0.03
passive	0.06	0.03	0.06	0.08	0.08	0.04	0.03	0.05	0.00	0.01	0.00	0.02	0.03	0.02	0.01	0.02

Table 29 Metrics for Various Models, New Movie 2

model	E				F				G				H			
answer	1		2		1		2		1		2		1		2	
metric	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
creativity	0.47	0.08	0.50	0.14	0.63	0.13	0.61	0.14	0.62	0.17	0.45	0.18	0.59	0.12	0.56	0.19
practicality	0.54	0.10	0.47	0.11	0.61	0.12	0.46	0.10	0.52	0.11	0.29	0.14	0.66	0.12	0.62	0.11
specificity	0.16	0.09	0.39	0.09	0.40	0.09	0.38	0.10	0.26	0.10	0.49	0.15	0.24	0.10	0.34	0.11
dialogue	0.42	0.05	0.39	0.09	0.36	0.20	0.59	0.09	0.39	0.17	0.43	0.19	0.33	0.15	0.65	0.10
thought	0.31	0.09	0.28	0.13	0.66	0.13	0.70	0.11	0.64	0.14	0.36	0.19	0.52	0.15	0.51	0.21
complexity	0.32	0.09	0.25	0.11	0.27	0.12	0.47	0.14	0.31	0.13	0.28	0.15	0.34	0.12	0.41	0.20
technicality	0.07	0.04	0.16	0.09	0.06	0.05	0.09	0.06	0.09	0.06	0.32	0.19	0.08	0.05	0.14	0.08
diversity	0.66	0.07	0.52	0.12	0.62	0.13	0.40	0.12	0.64	0.13	0.55	0.22	0.67	0.15	0.65	0.14
coherence	0.28	0.07	0.30	0.18	0.68	0.08	0.66	0.12	0.69	0.06	0.49	0.14	0.63	0.06	0.67	0.22
readability	0.86	0.05	0.73	0.06	0.66	0.06	0.49	0.06	0.65	0.09	0.62	0.08	0.76	0.09	0.59	0.18
named	0.13	0.06	0.28	0.09	0.23	0.06	0.25	0.07	0.16	0.09	0.52	0.19	0.17	0.07	0.33	0.11
lexical	0.35	0.08	0.35	0.13	0.66	0.10	0.64	0.11	0.71	0.09	0.49	0.13	0.73	0.10	0.84	0.09
avg. word	0.25	0.11	0.53	0.11	0.42	0.08	0.65	0.09	0.42	0.09	0.65	0.15	0.24	0.09	0.32	0.13
type token	0.54	0.05	0.56	0.10	0.51	0.10	0.35	0.10	0.57	0.14	0.57	0.26	0.55	0.12	0.73	0.12
dependency	0.05	0.02	0.04	0.02	0.10	0.05	0.13	0.03	0.08	0.07	0.08	0.06	0.06	0.04	0.11	0.09
passive	0.02	0.02	0.01	0.02	0.03	0.03	0.03	0.06	0.05	0.05	0.00	0.02	0.05	0.05	0.06	0.15