合肥工業大學
HEFEI UNIVERSITY OF TECHNOLOGY

The Best Proportion of Human and Generative AI During Collaboration: Insights from Synthetic Text Data

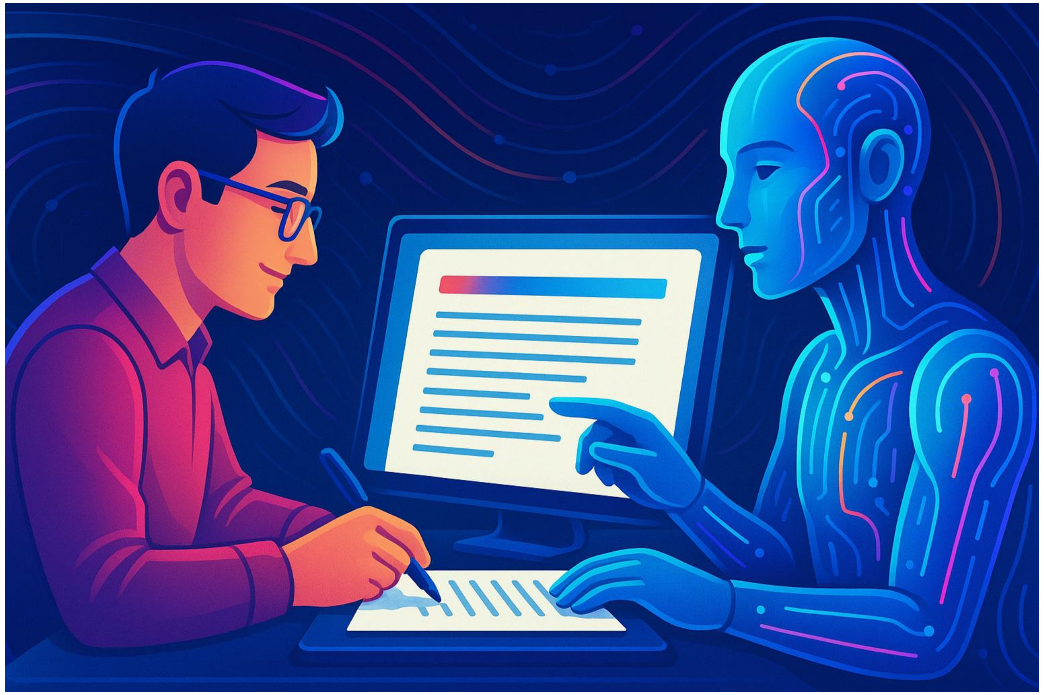
Charles Tian, Jun Pei

School of Management, Hefei University of Technology

Backgrounds

- As generative AI becomes increasingly powerful, its capabilities are comparable to or even surpass human performance in various tasks[1][2]. However, directly applying pure AI can lead to numerous issues[2][3].
- In certain fields and tasks, **human-AI collaboration** can produce content of higher quality than that created by humans alone[4], although in some cases, the quality may be inferior[5][6].

How can human-AI collaborative creation be structured to achieve higher content quality?

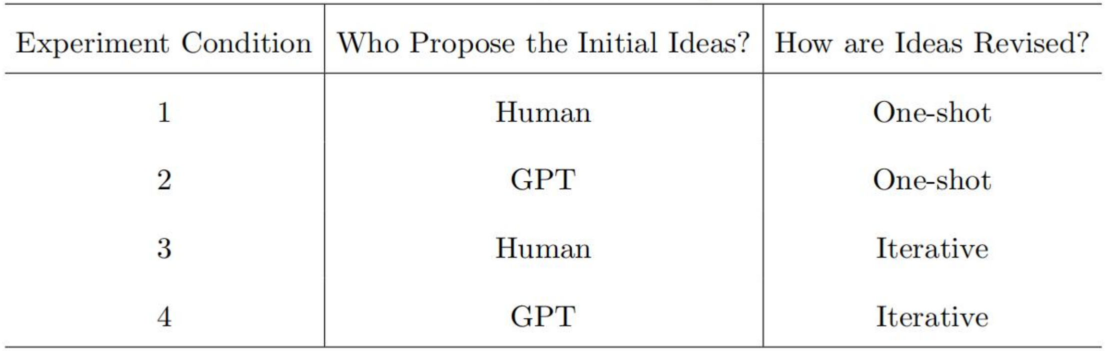


Backgrounds

Existing Works

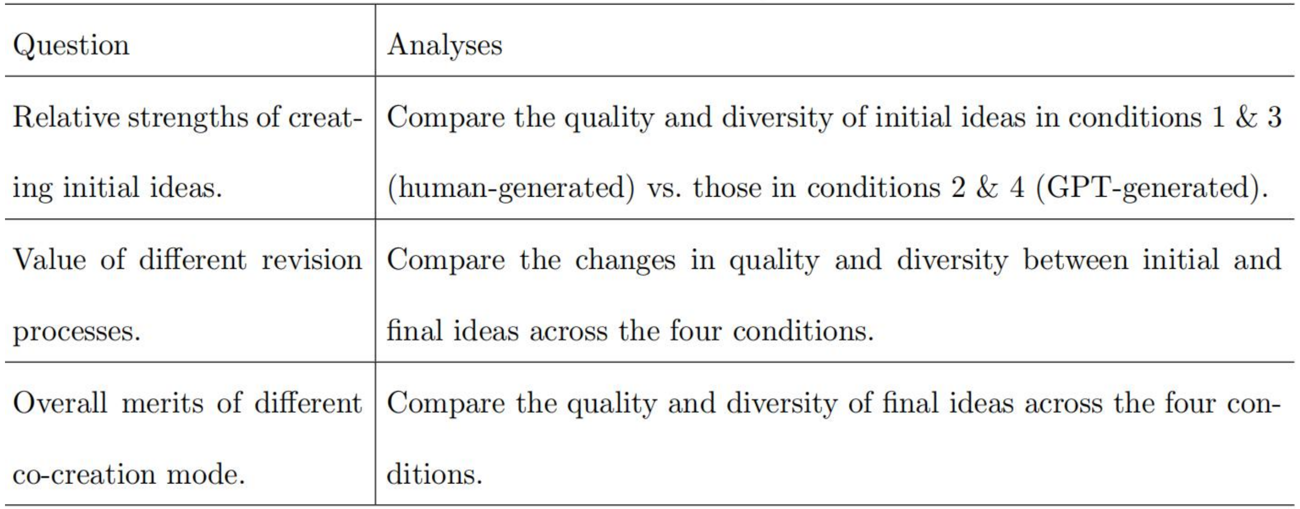
- Different AI usage methods[7][8]
- Certain tasks/abilities[9][10]
- Impacts[11]

Table 1 Experiment Design

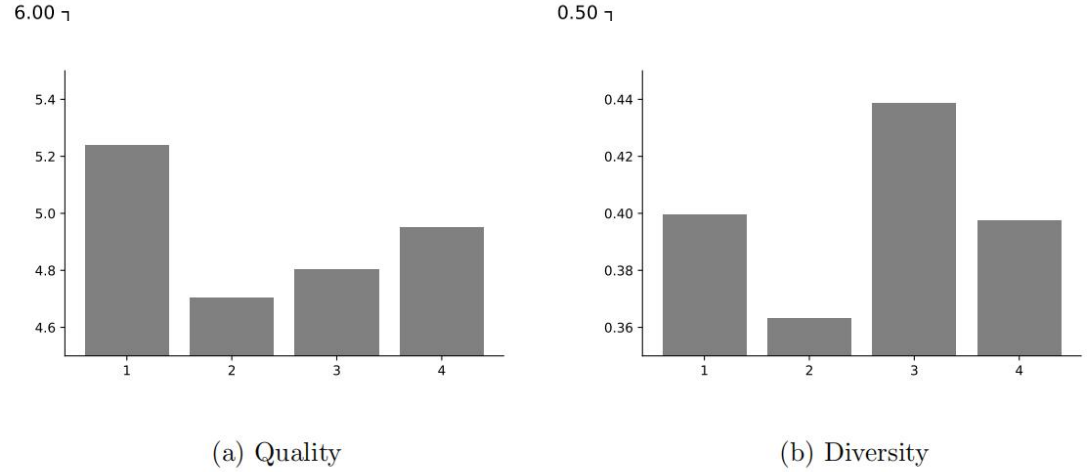


Experiment Condition	Who Propose the Initial Ideas?	How are Ideas Revised?
1	Human	One-shot
2	GPT	One-shot
3	Human	Iterative
4	GPT	Iterative

Table 2 Planned Analyses



Question	Analyses
Relative strengths of creating initial ideas.	Compare the quality and diversity of initial ideas in conditions 1 & 3 (human-generated) vs. those in conditions 2 & 4 (GPT-generated).
Value of different revision processes.	Compare the changes in quality and diversity between initial and final ideas across the four conditions.
Overall merits of different co-creation mode.	Compare the quality and diversity of final ideas across the four conditions.



(a) Quality

(b) Diversity

Fig **Figure 4 Overall Merits of Different Co-Creation Modes (Study 1)**

Figure 2 Relative Strengths of Creating Initial Ideas (Study 1)

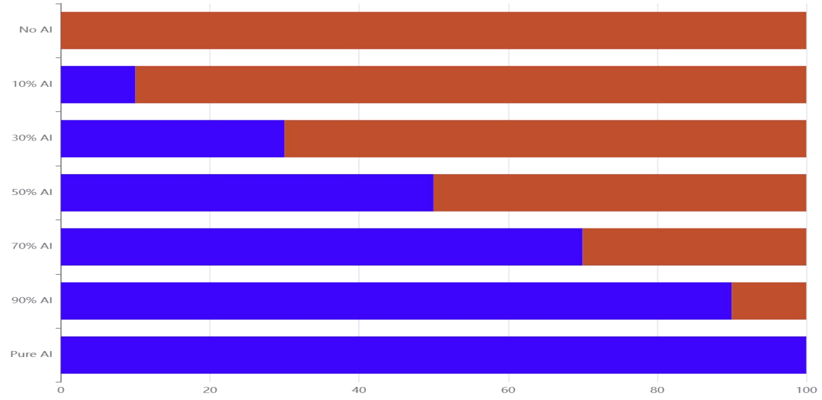
Motivations

Limitation:

Only on whether AI "participates," lacking finer granularity.

Motivation:

Can we investigate the proportion of AI ("how much" AI participates)?



Motivation: Varying levels of AI participation

Research Question:

What is the impact of varying levels of AI participation on the quality of human-AI collaborative generated content?

Takeaways:

Leveraging texts co-created by humans and AI based on different levels of AI involvement, we evaluate the content quality across groups.

The content quality is evaluated from five perspectives: content diversity, lexical richness, readability, linguistic acceptability, and emotional content.

Data

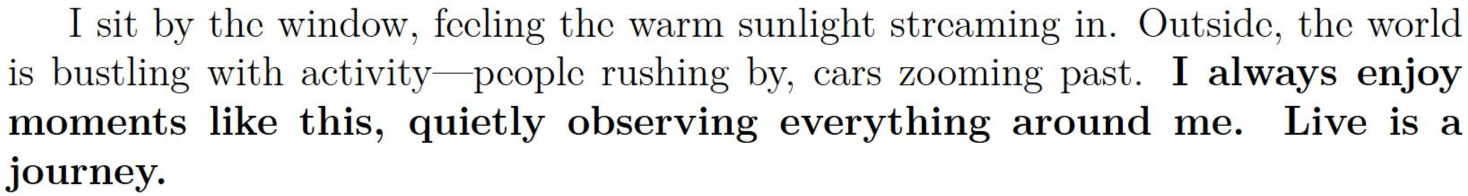
RQ1: Why choose text data?

- The most fundamental type of data.
- Its sequential nature facilitates grouping.

RQ2: How is AI participations defined?

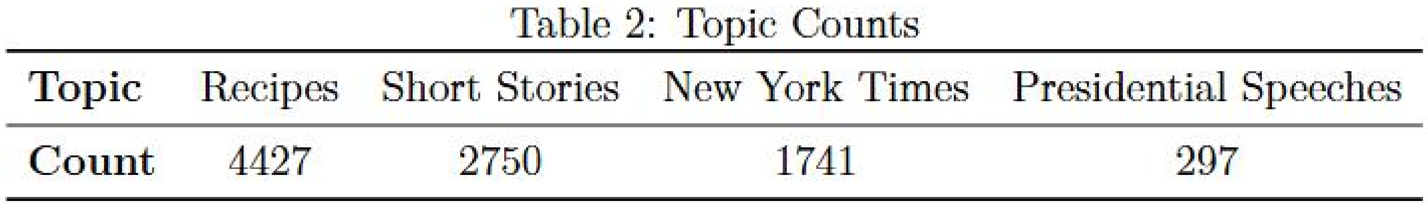
$$\text{AI engagement} = 1 - \frac{\text{bound} + 1}{\text{sentence}}$$

- Bound: The index of sentences at the boundary between human
- Sentence: The number of sentences in the document
- **RQ3: What is the basic situation of the data?**
- Source: [13]
- Basic information: Table 1, Table 2

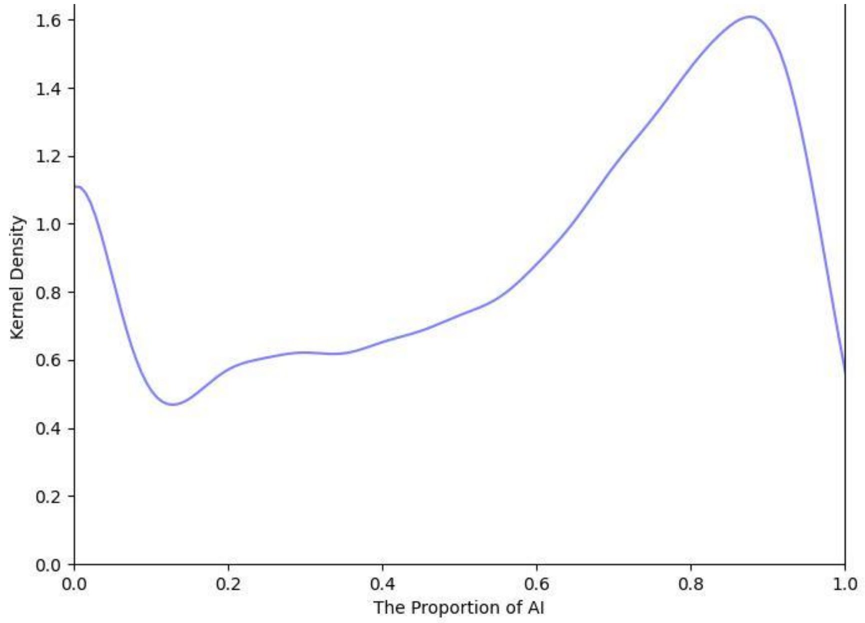


I sit by the window, feeling the warm sunlight streaming in. Outside, the world is bustling with activity—people rushing by, cars zooming past. **I always enjoy moments like this, quietly observing everything around me. Live is a journey.**

Table 2: Topic Counts



Topic	Recipes	Short Stories	New York Times	Presidential Speeches
Count	4427	2750	1741	297



AI engagement density map [14]

Methods

1 Content Diversity

- Using a pre-trained BERT model, the embedding for each document is obtained.
- The diversity metric is defined as follows [10]:

$$\text{Diversity} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{CosineDistance}(e_i, e_j)$$

- A higher diversity value indicates greater similarity in the content among the documents in the group, thereby suggesting lower content diversity.

Methods

2 Vocabulary richness

- The definition of the vocabulary richness metric is as follows:

$$\mathrm{TTR} = \frac{\mathrm{Types}}{\mathrm{Tokens}}$$

- Types: The number of unique vocabulary types in the document.
- Tokens: The total number of words in the document.
- A higher TTR (Type-Token Ratio) indicates fewer repeated words within the document, thereby reflecting greater vocabulary richness.

Methods

3 Readability

- The definition of the readability metric is as follows:

$$FRE = 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right)$$

- Total Words: The total number of words in the document
- Total Sentences: The total number of sentences in the document
- Total Syllables: The total number of syllables in the document
- A higher FRE (Flesch Reading Ease) score indicates that the average number of words per sentence and the average number of syllables per word are smaller, thereby suggesting higher readability of the document.

Methods

4 Language Acceptability

- Using a pre-trained BERT model that has been fine-tuned for logical judgment [15][16], the language acceptability of each sentence is obtained.
- The language acceptability metric is calculated as follows:

$$\mathrm{CoLA} = \frac{1}{sentence} \sum_{i=1}^{N} \mathrm{CoLA}_i$$

- A higher CoLA score indicates that more sentences in the document are considered linguistically acceptable, thereby suggesting greater language acceptability.

Methods

5 Sentiment Content

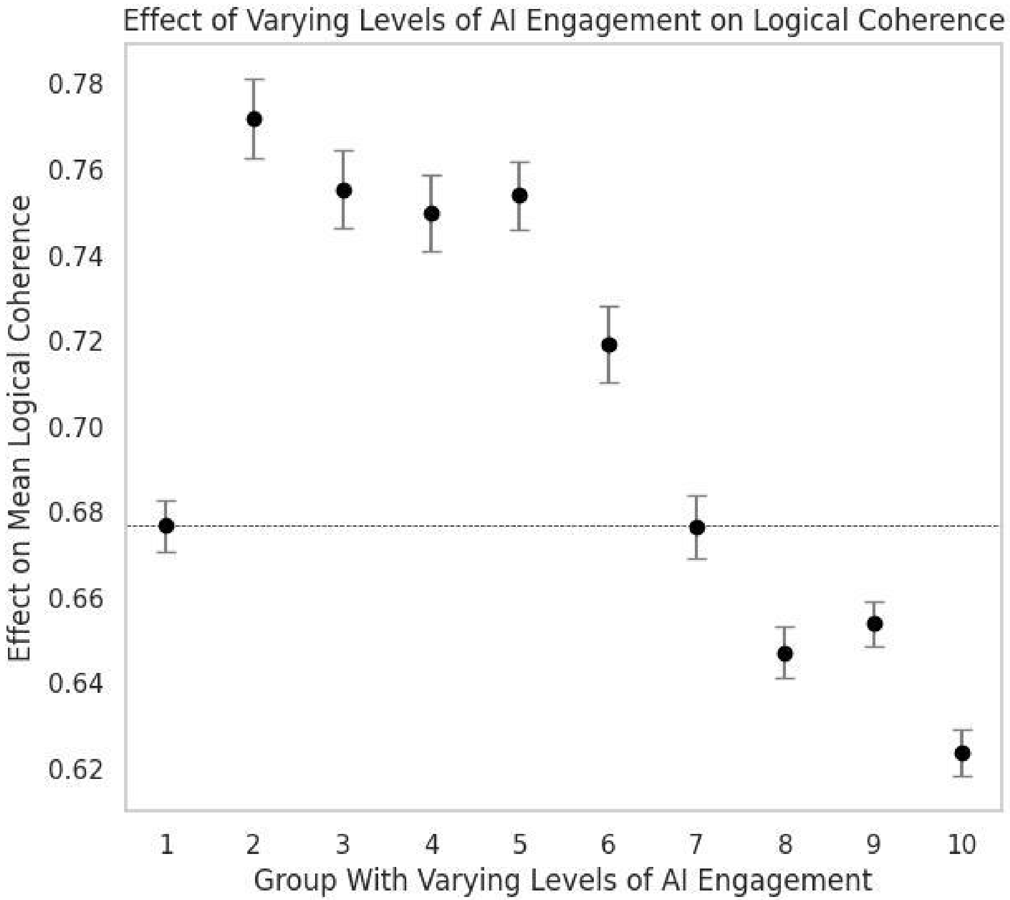
- Using a pre-trained Distil RoBERTa model that has been fine-tuned for sentiment analysis [17][18], the sentiment analysis results for each sentence are obtained.
- The sentiment content metric is calculated as follows:

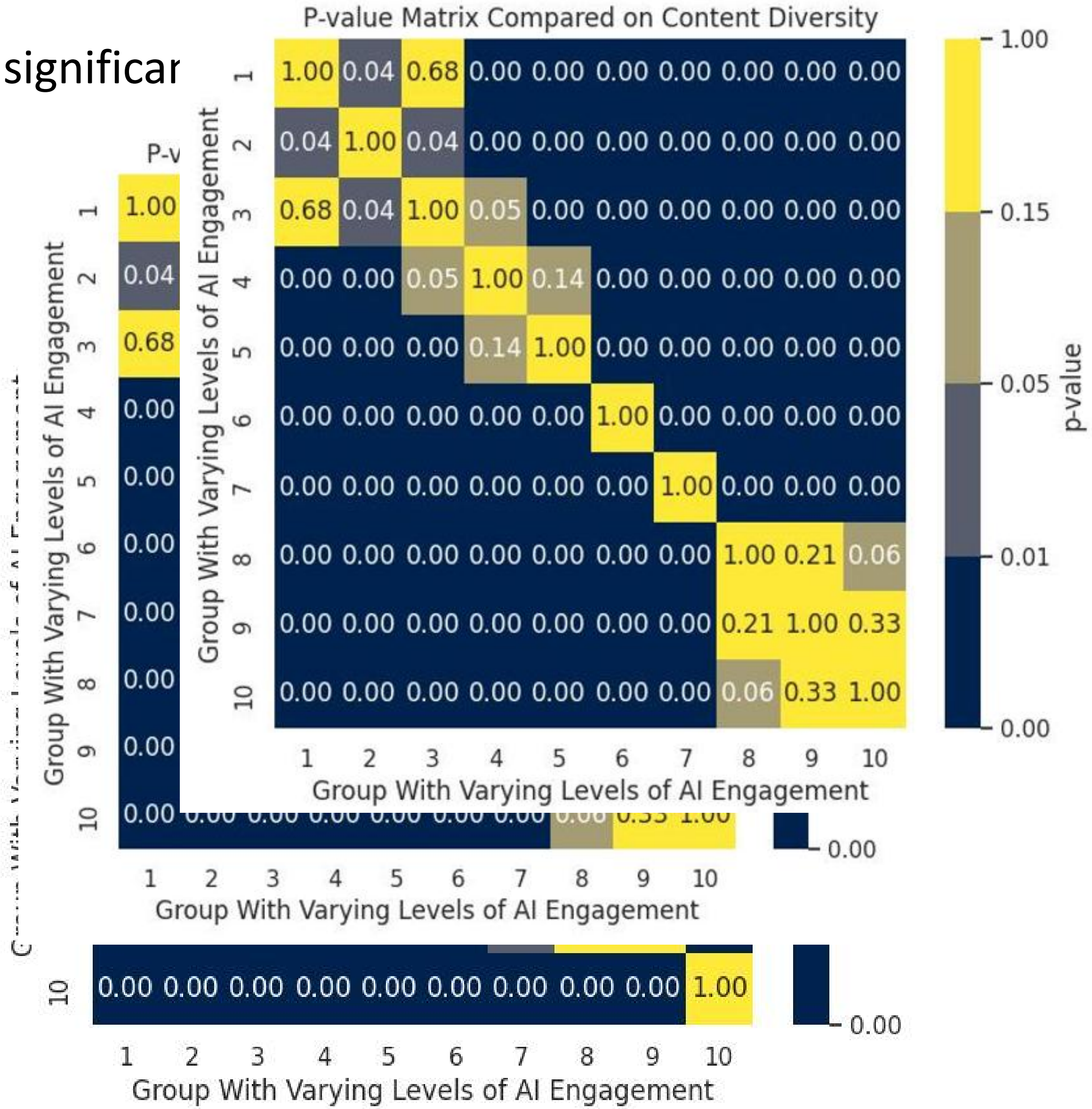
$$\text{sentiment} = \frac{1}{sentence} \sum_{i=1}^{N} \text{sentiment}_i$$

- A higher sentiment content score indicates that a larger proportion of the language in the document contains emotional elements, thereby suggesting greater emotional content.

Results

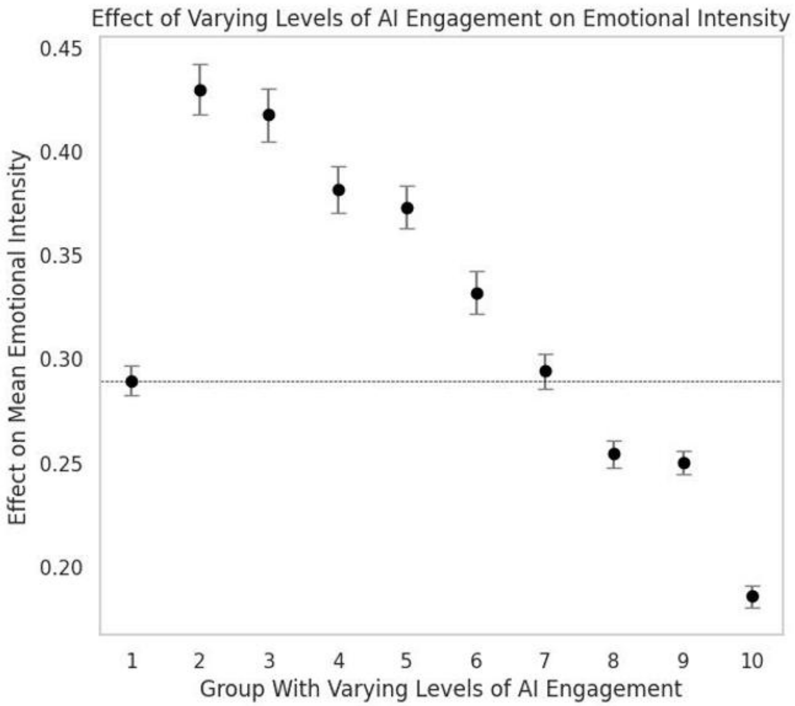
- Q1: Does varying levels of AI participation truly have a significar
- R1: Yes. Significance testing can demonstrate this.

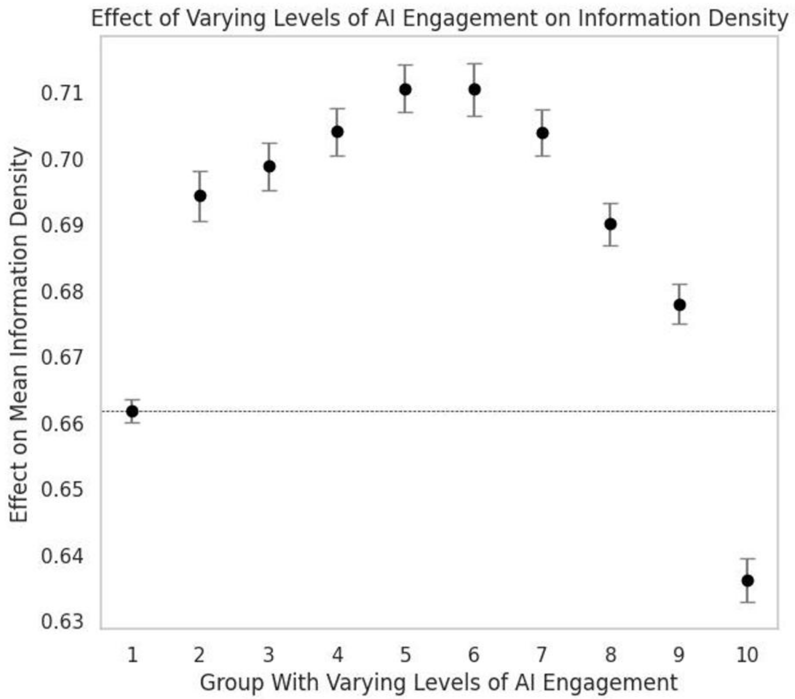


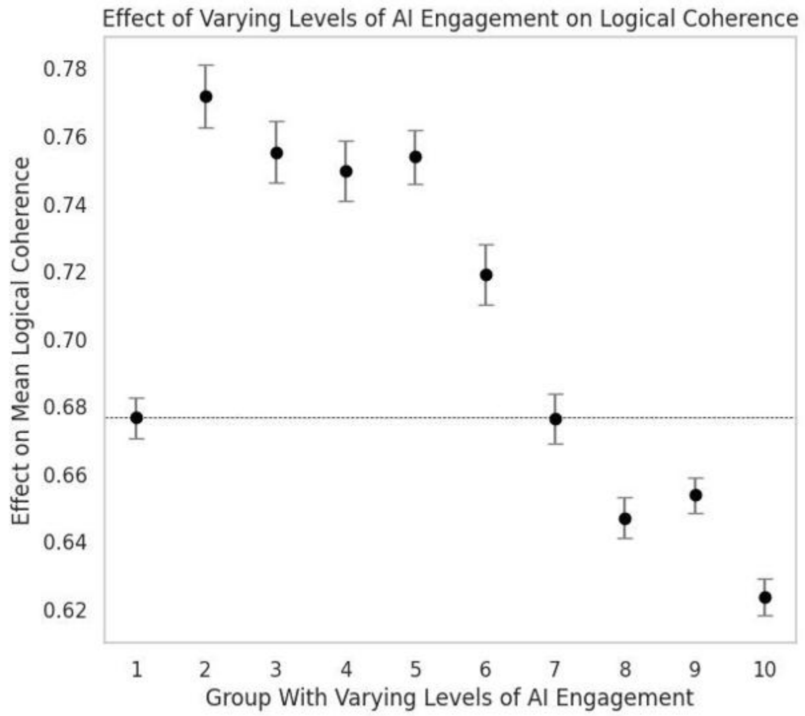


Results

- Q2: Is human-AI collaboration truly necessary?
- R2: Human-AI collaboration is more effective than human-led approaches, which in turn outperform AI-dominant methods. Additionally, performance without human involvement is significantly lower than that with human participation.

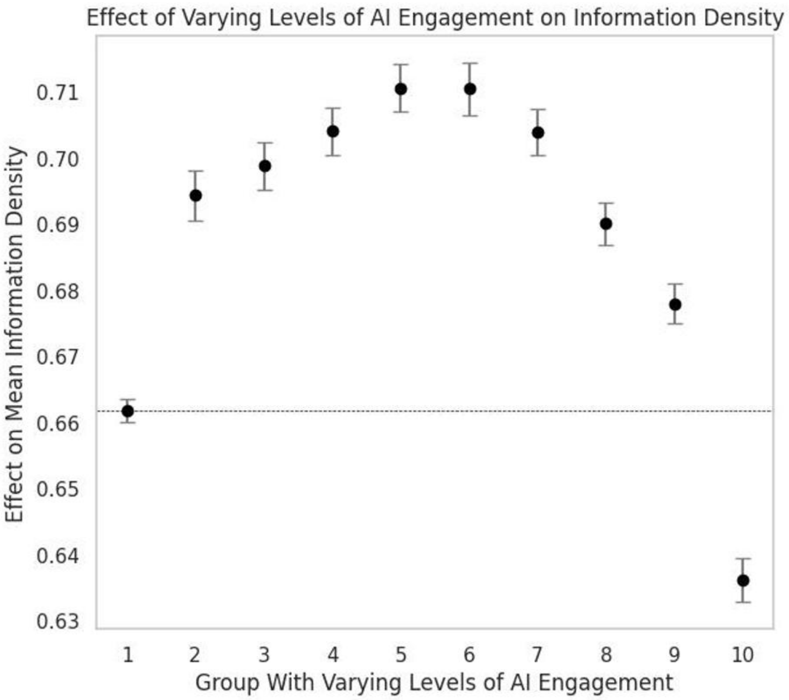

Effect of Varying Levels of AI Engagement on Emotional Intensity

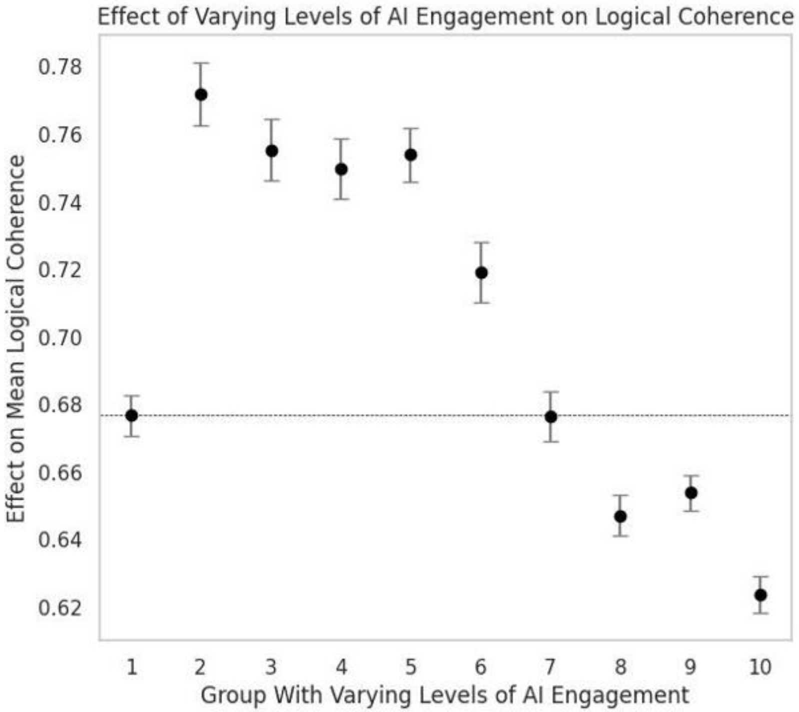

Effect of Varying Levels of AI Engagement on Information Density

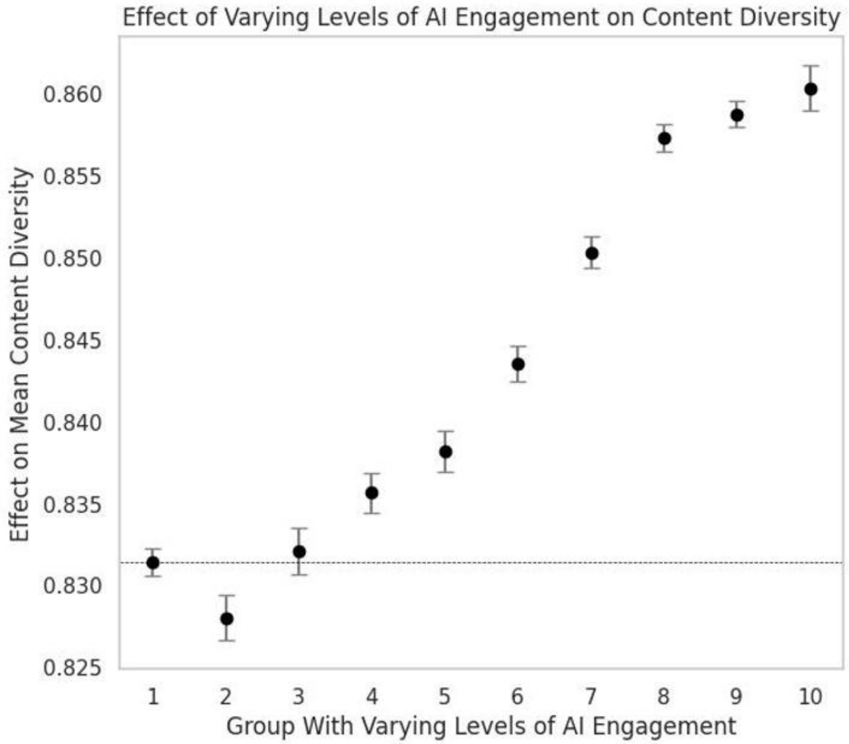

Effect of Varying Levels of AI Engagement on Logical Coherence

Results

- Q3: How does the overall text quality change as AI participation increases?
- R3: Human-dominant > Pure human / AI-dominant > Pure AI[7]. It initially increases and then decreases. When AI usage is relatively low, performance is generally better. After surpassing a certain threshold of AI involvement, performance declines sharply.

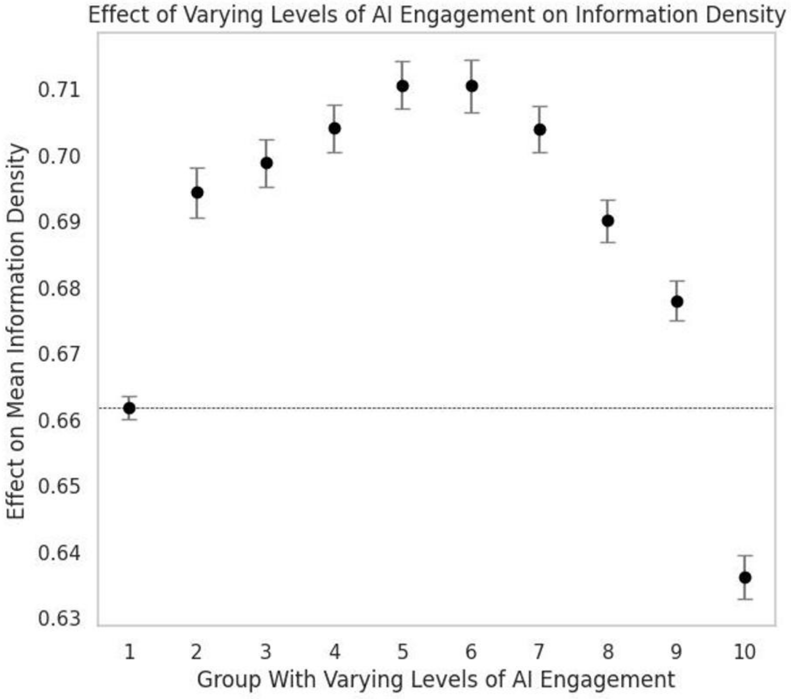


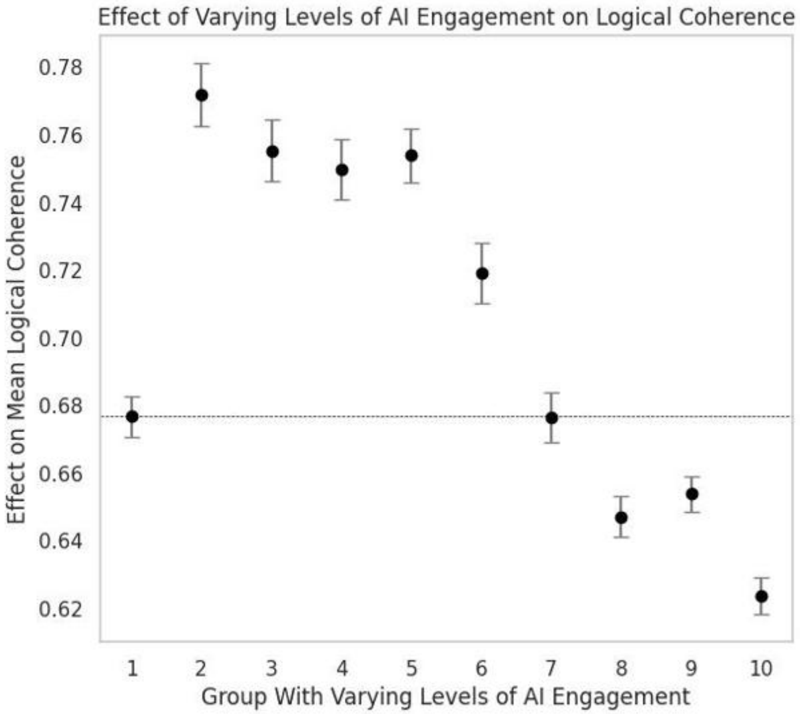


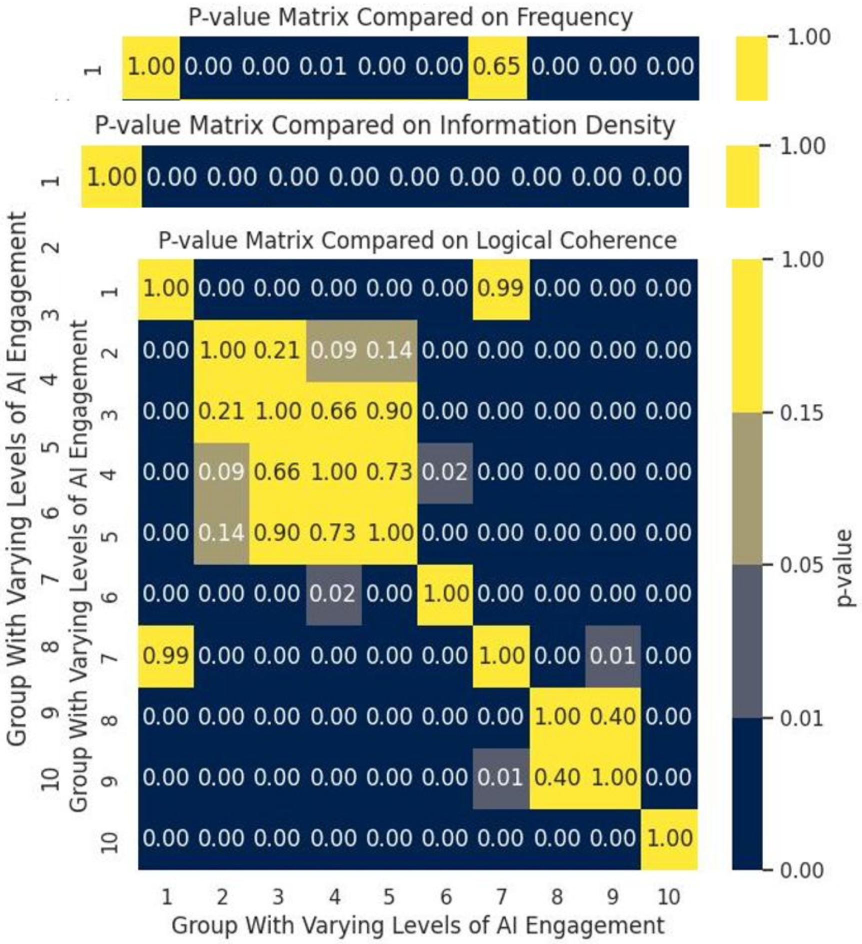


Results

- Q4: What is the optimal ratio of human to AI involvement?
- R4: It is difficult to draw a definitive conclusion; performance with a human-dominant ratio of 20% to 50% is generally similar.







P-value Matrix Compared on Frequency

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.00	0.01	0.00	0.00	0.65	0.00	0.00	0.00

P-value Matrix Compared on Information Density

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

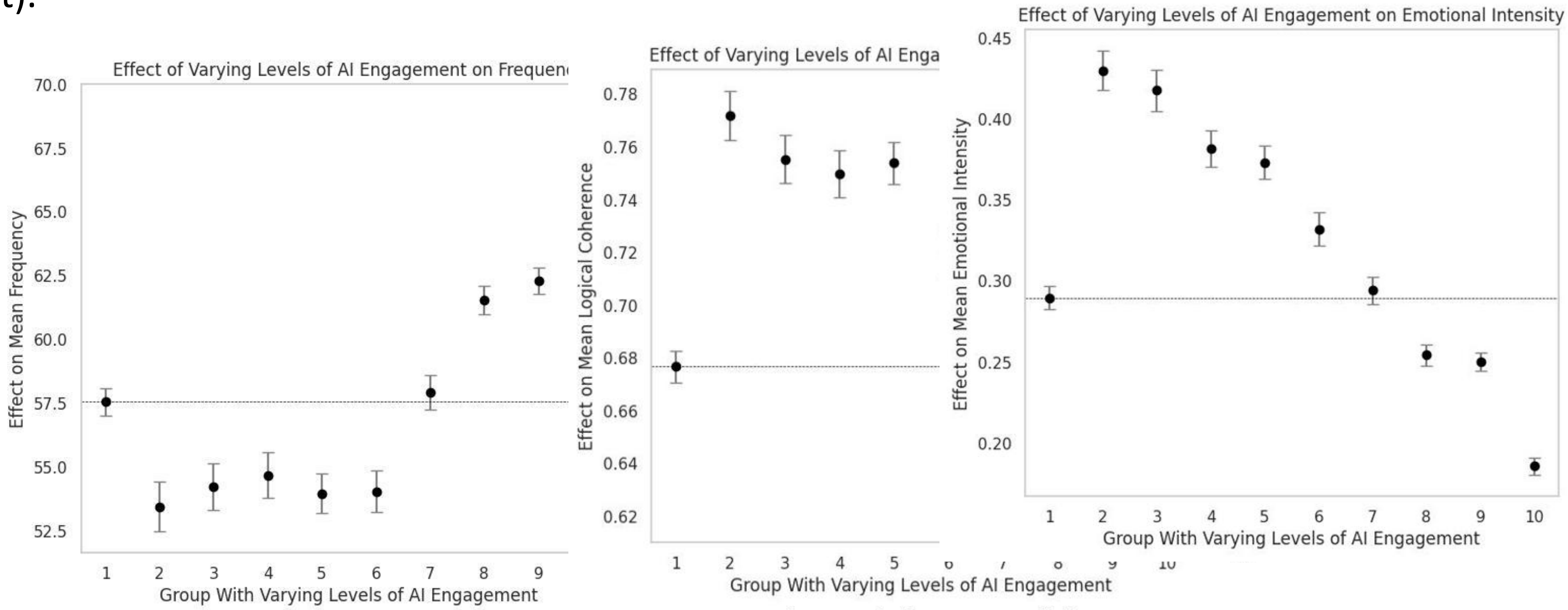
P-value Matrix Compared on Logical Coherence

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00
2	0.00	1.00	0.21	0.09	0.14	0.00	0.00	0.00	0.00	0.00
3	0.00	0.21	1.00	0.66	0.90	0.00	0.00	0.00	0.00	0.00
4	0.00	0.09	0.66	1.00	0.73	0.02	0.00	0.00	0.00	0.00
5	0.00	0.14	0.90	0.73	1.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.02	0.00	1.00	0.00	0.00	0.00	0.00
7	0.99	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.40	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.40	1.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Results

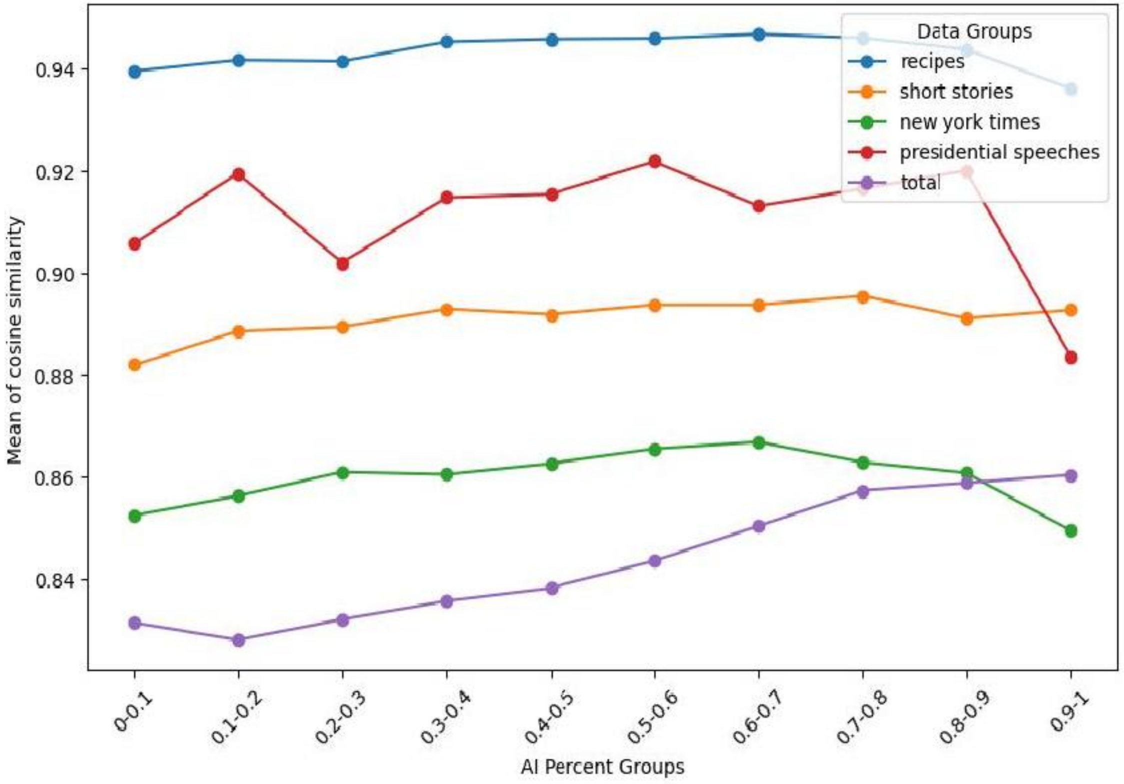
Exceptions in Readability Metrics

- As AI participation increases, readability first decreases and then increases, so the order is AI-dominant / human-dominant.
- Speculation: AI may be better at generating content with higher consistency (variance: AI-dominant > human-dominant).

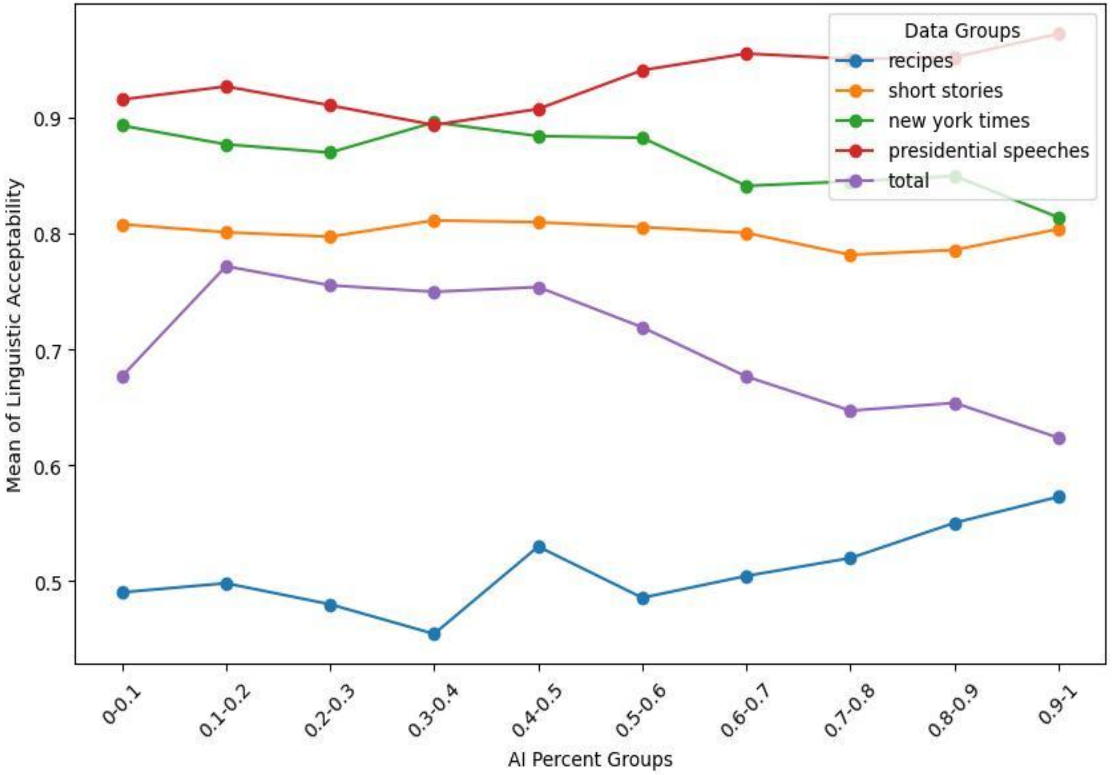


Results

- Q5: Do the documents for each topic follow this pattern?
- R5: There are significant differences between various topics — further explanation of the reasons why different studies have reached conflicting conclusions.



Trends of content diversity metrics across documents of different topics



Trends of language acceptability metrics across documents of different topics

Conclusions

Key Insights:

- Different proportions of generative AI significantly impact the quality of generated content.
- In most cases, human-AI collaboration outperforms pure human effort, which in turn outperforms pure AI; therefore, it is encouraged to use generative AI.
- It is also advised to think independently and limit reliance on generative AI, as excessive use can lead to decreased quality, sometimes even worse than human-generated content.
- The impact of generative AI varies greatly across different fields, so specific analysis is needed for each particular case.

Limitations:

- This work only shows a special case of collaboration, which means that human-write, GenAI-complement, which may can’t generalize to active collaboration.
- Evaluation metrics is not comprehensive and typical. Pure automatic evaluation but not human evaluation or llm-as-a-judge, which may not show the precise results.

Selected References

[1] Schreiner M (2023) Gpt-4 beats nearly 2,700 students in creative thinking. The Decoder URL https://the-decoder.com/gpt-4-beats-nearly-2700-students-in-creative-thinking/.

[2] Cheng L, Li X, Bing L (2023) Is gpt-4 a good data analyst? arXiv preprint arXiv:2305.15038.

[3] Wach, K., Duong, C.D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J.,& Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2), 7-24.

[4] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654):187–192, July 2023. doi: 10.1126/science.adh2586.

[5] Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022, June). News from generative artificial intelligence is believed less. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 97-106).

[6] Vaccaro, M., Almaatouq, A., & Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis[J]. Nature Human Behaviour, 2024: 1-11

[7] Lu, T., & Zhang, Y. (2024). 1+ 1> 2? information, humans, and machines. Information Systems Research.

[8] Chen Z, Chan J (2023) Large language model in creative work: The role of collaboration modality and user expertise. Available at SSRN 4575598

[9] Zhou, E., & Lee, D. (2024). Generative ai, human creativity, and art. Eric Zhou, Dokyun Lee, Generative artificial intelligence, human creativity, and art, PNAS Nexus, 3(3).

[10] Wang, W., Yang, M., & Sun, T. (2023). Human-AI co-creation in product ideation: The dual view of quality and diversity. Available at SSRN 4668241.

[11] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models, March 2023.

[12] Padmakumar V, He H (2023) Does writing with language models reduce content diversity? arXiv preprint arXiv:2309.05196 .

[13] Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023, June). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 11, pp. 12763-12771).

[14] Zhao, W., Wang, W., & Viswanathan, S. (2024). Spillover Effects of Generative AI on Human-Generated Content Creation: Evidence from a Crowd-Sourcing Design Platform. Available at SSRN 4693181.

[15] Üyük, C., Rovó, D., Kolli, S., Varol, R., Groh, G., & Dementieva, D. (2024). Crafting Tomorrow's Headlines: Neural News Generation and Detection in English, Turkish, Hungarian, and Persian. arXiv preprint arXiv:2408.10724.

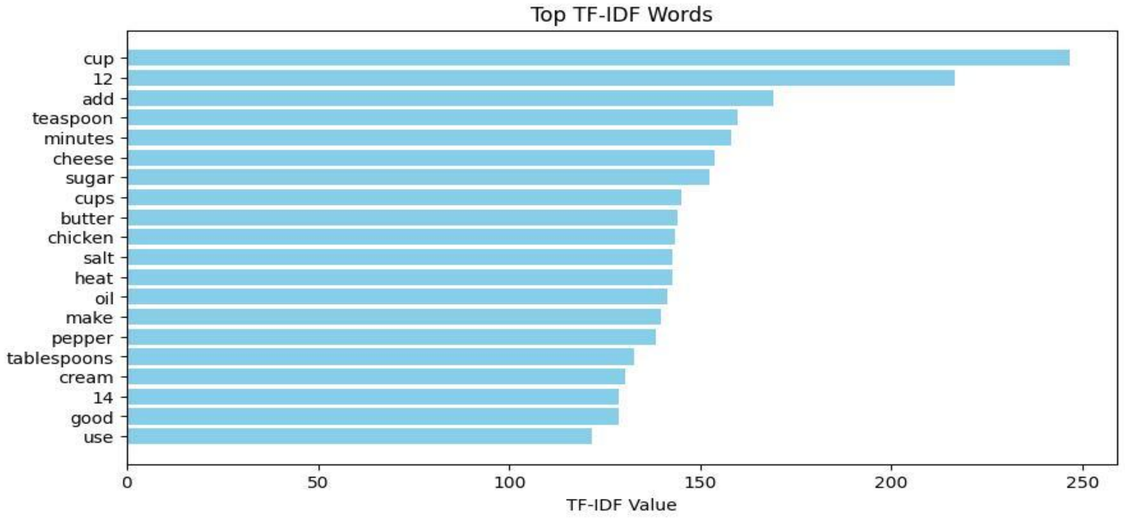
[16] https://huggingface.co/textattack/bert-base-uncased-CoLA

[17] Wang, W., & Li, B. (2024). Learning Personalized Privacy Preference From Public Data. Information Systems Research.

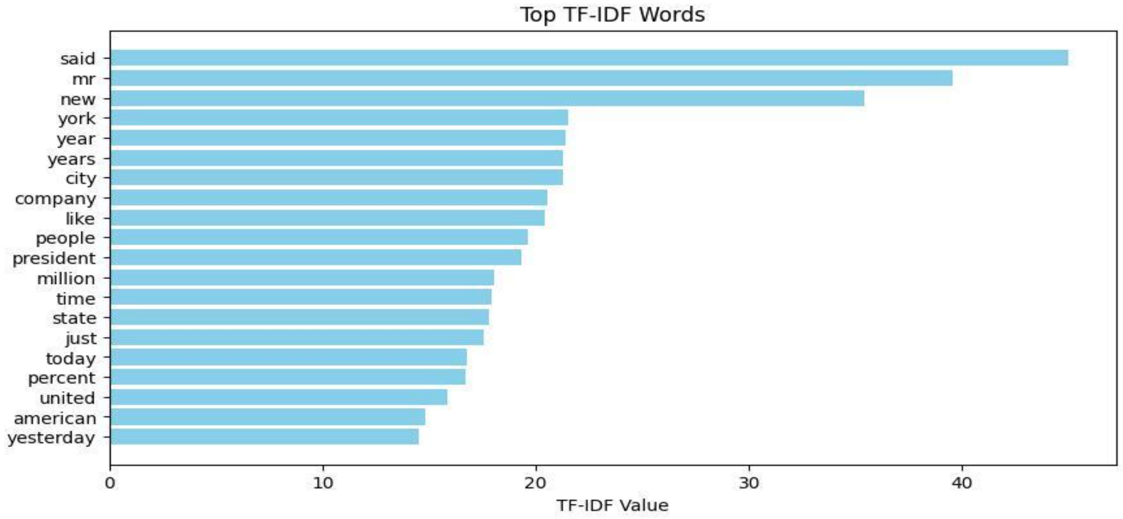
[18] Hartmann J (2022) Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Appendix A: Basis of Data

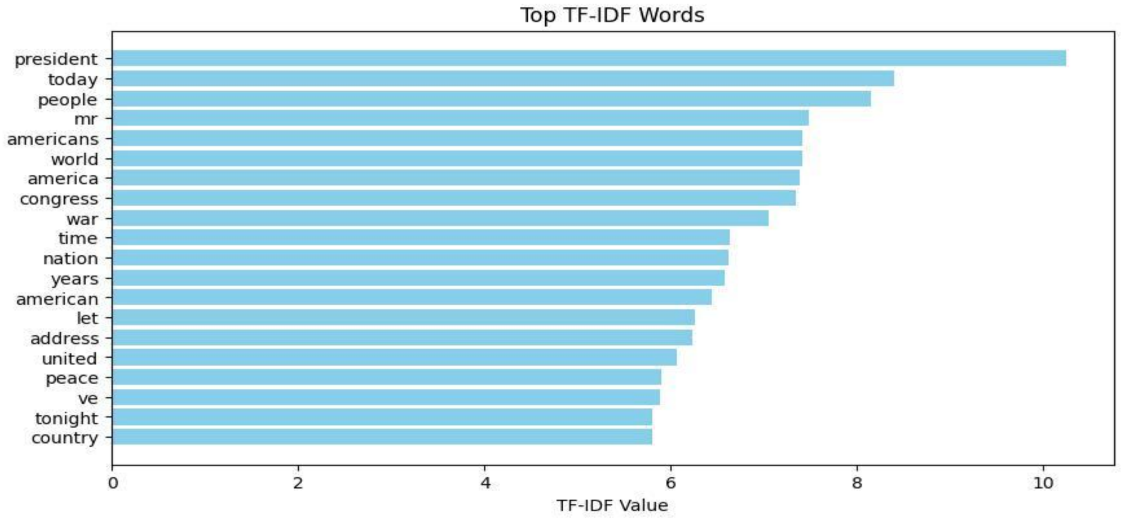
TF-IDF analysis results for recipes



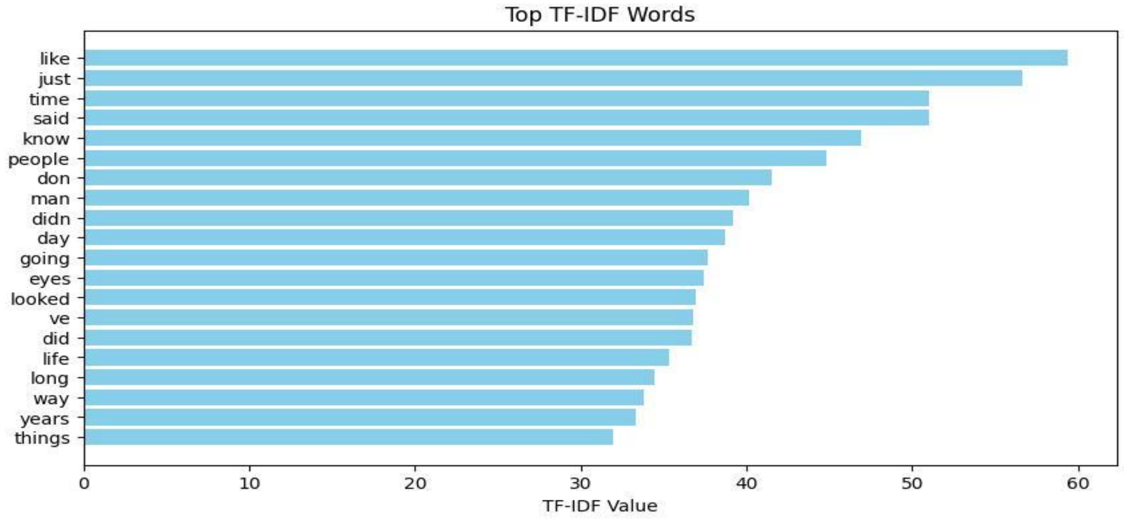
TF-IDF analysis results for New York Times news articles



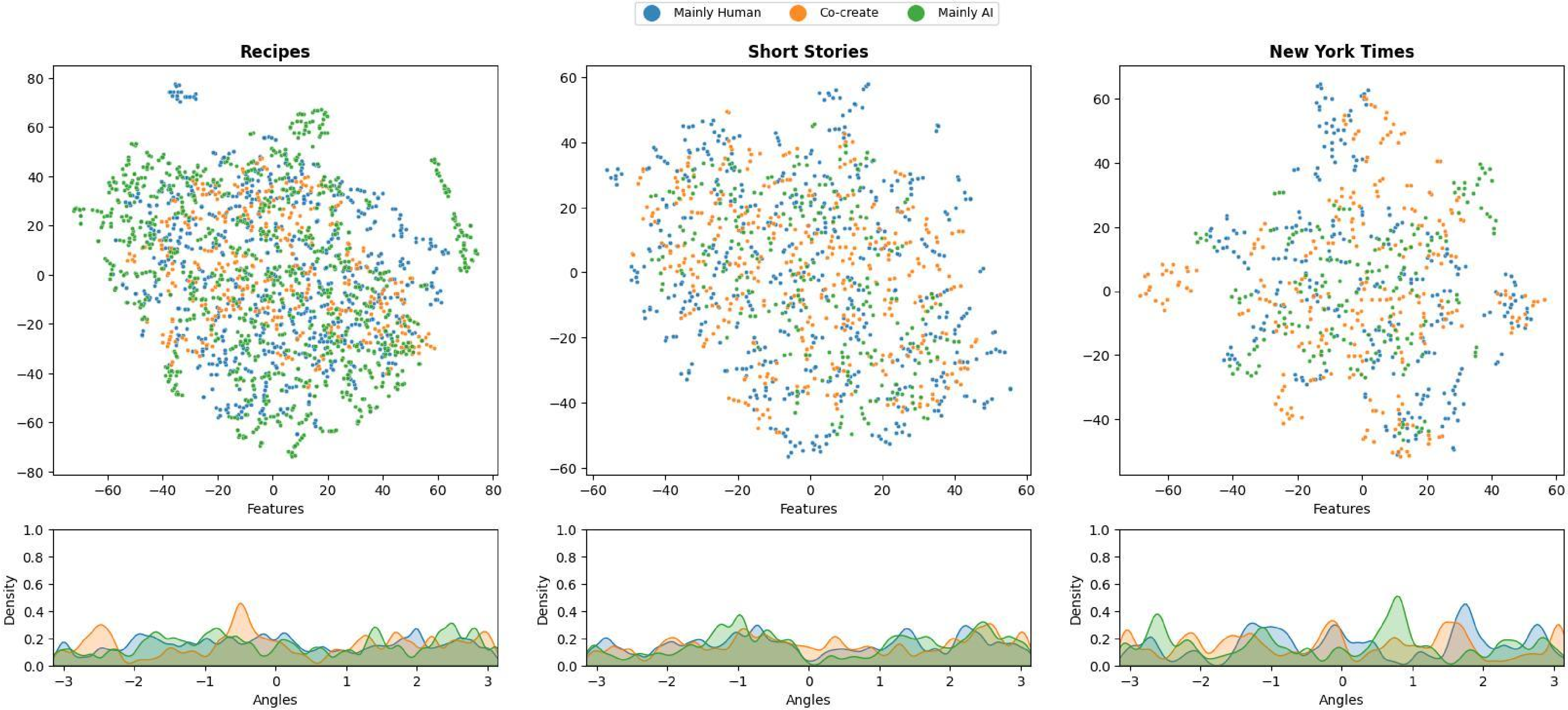
TF-IDF analysis results for presidential speeches



TF-IDF analysis results for short stories



Appendix B: Embedding Smoothness Analysis


Mainly Human
Co-create
Mainly AI
Recipes
Short Stories
New York Times
Features
Angles
Density