Face4Rag: Factual Consistency Evaluation for Retrieval Augmented Generation in Chinese

Yunqi Xu, Tianchi Cai, Jiyan Jiang, and Xierui Song. 2024. Face4Rag: Factual Consistency Evaluation for Retrieval Augmented Generation in Chinese. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 6083–6094. https://doi.org/10.1145/3637528.3671656

Presenter: Cheng Jhe Lee

National Cheng Kung University

國立成功大學
National Cheng Kung University

Outline

- Introduction
- Related Works
- Methodology
  - Face4RAG Benchmark
  - Logic-Enhanced Factual Consistency Evaluation
- Experiment
- Conclusion
- References

Outline

- Introduction
- Related Works
- Methodology
  - Face4RAG Benchmark
  - Logic-Enhanced Factual Consistency Evaluation
- Experiment
- Conclusion
- References

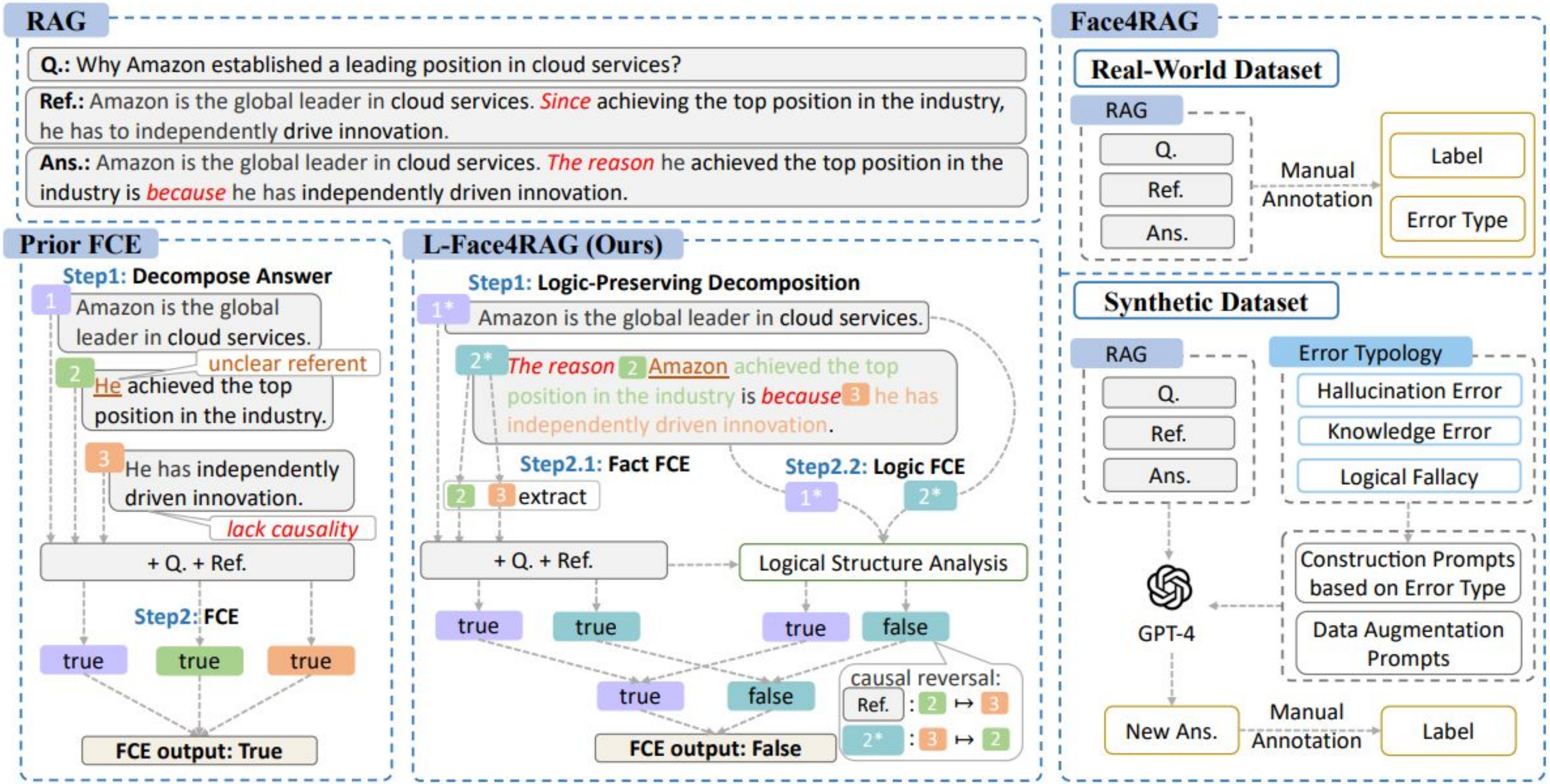
Introduction

- **Background**
  - Retrieval Augmented Generation (RAG): enhances the context of LLMs with relevant passages
  - Passages retrieved from external sources like retrievers or search engines
  - RAG shows strong performance on knowledge-intensive tasks
  - Key applications: open domain conversation and question answering

Introduction

- **Background**
  - Leading RAG systems like Bing Chat and Perplexity: only slightly over 50% factual consistency
  - Many Factual Consistency Evaluation (FCE) methods in RAG tasks evaluated on datasets from specific LLMs
  - Lack of comprehensive benchmarks for testing FCE methods on different LLMs
  - FCE methods may fail to detect error types from other LLMs

Introduction



Outline

Related Works

- **Traditional FCE Methods**
  - Evaluating the factuality of model generated results is widely studied across various language model generation domains, like
    - text summarization [15]
    - dialogue summary [47]
    - question-answering [4].

[4] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. arXiv preprint arXiv:2306.04181 (2023).

[15] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554 (2023).

[47] Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization. arXiv preprint arXiv:2305.16548 (2023).

Related Works

- **FCE for Long-form Answers**
  - To effectively evaluate factuality of long answers, recent FCE research mostly take a two step approaches [23, 26]
    - In the first step, the long-form answer is decomposed into shorter segments [10, 18, 23, 24, 26, 31, 18]
    - The second step evaluates the verifiability of each segment with respect to the given reference text [25, 26, 45], which can be efficiently done by modern general purpose LLM [9, 31], e.g., GPT4.

Related Works

- **FCE Benchmarks**
  - Prior benchmarks for FCE mostly focus on specialized tasks like summarization [13, 24, 39]. For FCE in RAG, existing benchmarks are derived from specific LLMs, such as Refchecker [18] and FELM [9], which are constrained by the error type distribution of the underlying LLMs.

[9] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. arXiv preprint arXiv:2310.00741 (2023).

[13] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217 (2023).

[18] Xiangkun Hu, Dongyu Ru, Qipeng Guo, Lin Qiu, and Zheng Zhang. 2023. RefChecker for Fine-grained Hallucination Detection. (2023). https://github.com/ amazon-science/RefChecker

[24] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840 (2019).

[39] Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. arXiv preprint arXiv:2205.12854 (2022).

Outline

- Introduction
- Related Works
- Methodology
  - Face4RAG Benchmark
  - Logic-Enhanced Factual Consistency Evaluation
- Experiment
- Conclusion
- References

Methodology - Face4RAG Benchmark

- **Introduction**
  - Error-type-oriented synthetic dataset
    - includes nine types of errors belonging to three main categories
  - A real-world dataset
    - constructed from six commonly used LLMs

Methodology - Face4RAG Benchmark

- **Error Typology in FCE**
  - Hallucination Error
    - Hallucination Error (Hallu.)
  - Knowledge Error
    - Contradiction Error (KCont.)
    - Entity Inversion Error (KInve.)
    - Conflation Error (KConf.)
    - Conceptual Substitution Error (KConc.)

Methodology - Face4RAG Benchmark

Error Type	Original Text	Factual Inconsistent Text
KCont.	功能饮料中的维生素、矿物质等，对于运动后快速补充身体营养，消除疲劳具有一定作用。 The vitamins and minerals in energy drinks play a certain role in quickly replenishing nutrients and eliminating fatigue after exercise.	功能饮料中的元素、微生物等，对于运动后快速补充身体营养，增加疲劳具有一定作用。 The vitamins and minerals in energy drinks play a certain role in quickly replenishing nutrients and inducing fatigue after exercise.
KInve.	一般蚕可以活一个多月，其中从孵化到结茧根据季节不同大约是25-32天，变成蛹后有15-18天，最后成蛾是1-3 天。 A typical silkworm can live for just over a month, during which the period from hatching to cocooning varies roughly from 25 to 32 days depending on the season, followed by 15 to 18 days as a pupa, and finally 1 to 3 days as a moth.	一般蚕可以活一个多月，其中从孵化到结茧根据季节不同大约是15-18天，变成蛹后有25-32天，最后成蛾是1-3 天。 A typical silkworm can live for just over a month, during which the period from hatching to cocooning varies roughly from 15 to 18 days depending on the season, followed by 25 to 32 days as a pupa, and finally 1 to 3 days as a moth.
KConf.	防晒霜中的无机化学物质可以反射或散射皮肤上的光线，而有机（碳基）化学物质可以吸收紫外线。 The inorganic chemicals in sunscreen can reflect or scatter light on the skin, while organic (carbon-based) chemicals can absorb ultraviolet rays.	防晒霜中的无机化学物质和有机（碳基）化学物质都可以反射或散射皮肤上的光线、吸收紫外线。 Both the inorganic chemicals and organic (carbon-based) chemicals in sunscreen can reflect or scatter light on the skin and absorb ultraviolet rays.
KConc.	随着健康意识的增强，越来越多的人开始注重膳食平衡。 With the increasing awareness of health, more and more people are beginning to focus on a balanced diet.	随着健康意识的增强，越来越多的人开始注重膳食的有机质量。 With the increasing awareness of health, more and more people are beginning to focus on the organic quality of their diets.

Methodology - Face4RAG Benchmark

- **Error Typology in FCE**
  - Logical Fallacy
    - Overgeneralization Error (LOver.)
    - Causal Confusion Error (LCaus.)
    - Confusing Sufficient and Necessary Conditions Error (LConf.)
    - Inclusion Relation Error (LIncl.)
    - Other Logical Fallacy (LOthe.)

Methodology - Face4RAG Benchmark

Error Type	Original Text	Factual Inconsistent Text
LOver.	一般的我们平时见到的蜘蛛都是晚上出来。 The spiders that we usually see tend to come out at night.	一般的我们平时见到的昆虫都是晚上出来。 The insects that we usually see tend to come out at night.
LCaus.	随着信息技术的快速发展，大数据在各行各业中的应用越来越广泛。 With the rapid development of information technology, the application of big data across various industries is becoming increasingly widespread.	大数据在各行各业中的应用越来越广泛，这导致了信息技术的快速发展。 The application of big data across various industries is becoming increasingly widespread, leading to the rapid development of information technology.
LConf.	为了获得某项荣誉学生奖学金，学生必须具备以下条件：成绩优秀、品行端正、参加社会实践活动。 To receive a certain honor student scholarship, students must meet the following criteria: excellent academic performance, good moral character, and participation in social practice activities.	学生成绩优秀、品行端正就可以获得某项荣誉学生奖学金。 Students with excellent academic performance and good moral character can receive a certain honorary student scholarship.
LIncl.	坚持锻炼身体可以提高心肺能力，加强肌肉的耐力，提高身体的抗疲劳能力。 Regular exercise can enhance cardiorespiratory fitness, strengthen muscle endurance, and improve the body's resistance to fatigue.	坚持锻炼身体可以提高心肺能力，例如加强肌肉的耐力、提高身体的抗疲劳能力。 Regular exercise can enhance cardiorespiratory fitness, such as strengthening muscle endurance and improving the body's resistance to fatigue.

Methodology - Face4RAG Benchmark

- **Synthetic Dataset**
  - Based on WebCPM [36]
  - Negative Samples
  - Positive Samples
  - Human Annotation Refinement

[36] Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. WebCPM: Interactive Web Search for Chinese Long-form Question Answering. arXiv preprint arXiv:2305.06849 (2023).

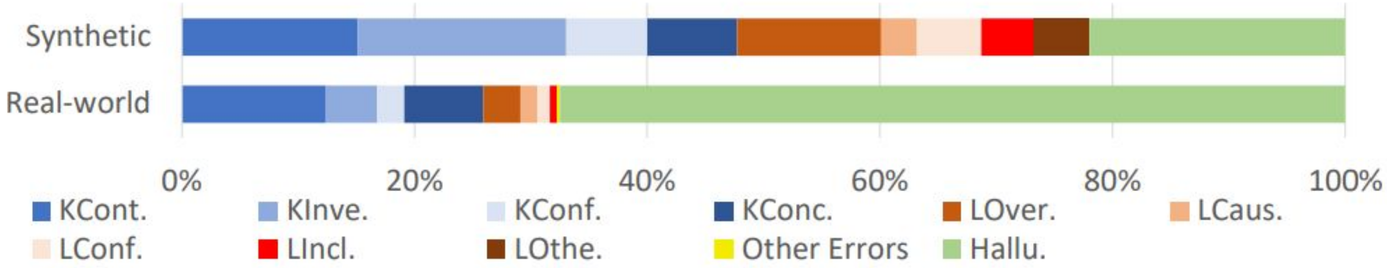
Methodology - Face4RAG Benchmark

- **Real-World Dataset**
  - gpt-3.5-turbo (GPT-3.5) [33]
  - gpt-4-turbo (GPT-4) [2]
  - Baichuan2- 13B-Chat (Baichuan2) [5]
  - ChatGLM3 [44]
  - Qwen-14B-Chat (Qwen) [3]
  - Chinese-Alpaca-2-13B-16k(Alpaca2 (CH)) [11]

Methodology - Face4RAG Benchmark

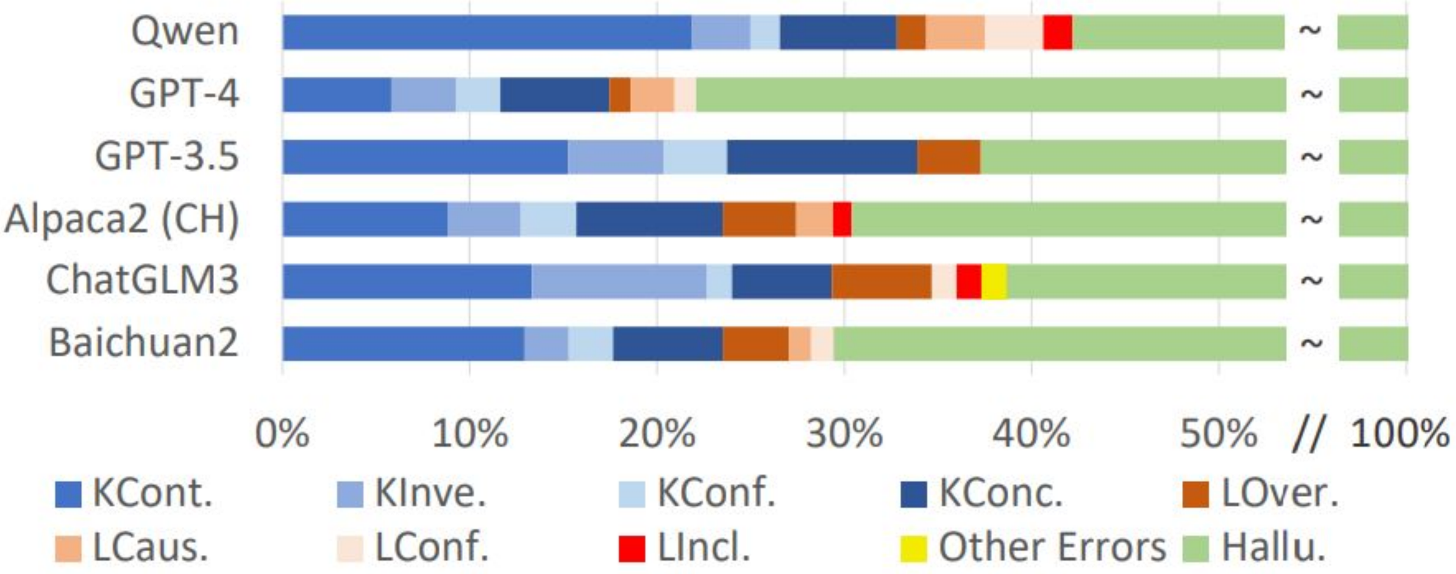
- **Overall Error Type Distribution**

Statistics	Synthetic Dataset		Real-world Dataset	
	Answer	Segment	Answer	Segment
Num. Samples	1299	6737	1200	6143
Avg. Length	289.3	45.4	307.7	45.2
Positive Rate	30.3%	55.8%	63.3%	85.6%



Methodology - Face4RAG Benchmark

- **Real-World Dataset - Error Distribution of Various Models**



Outline

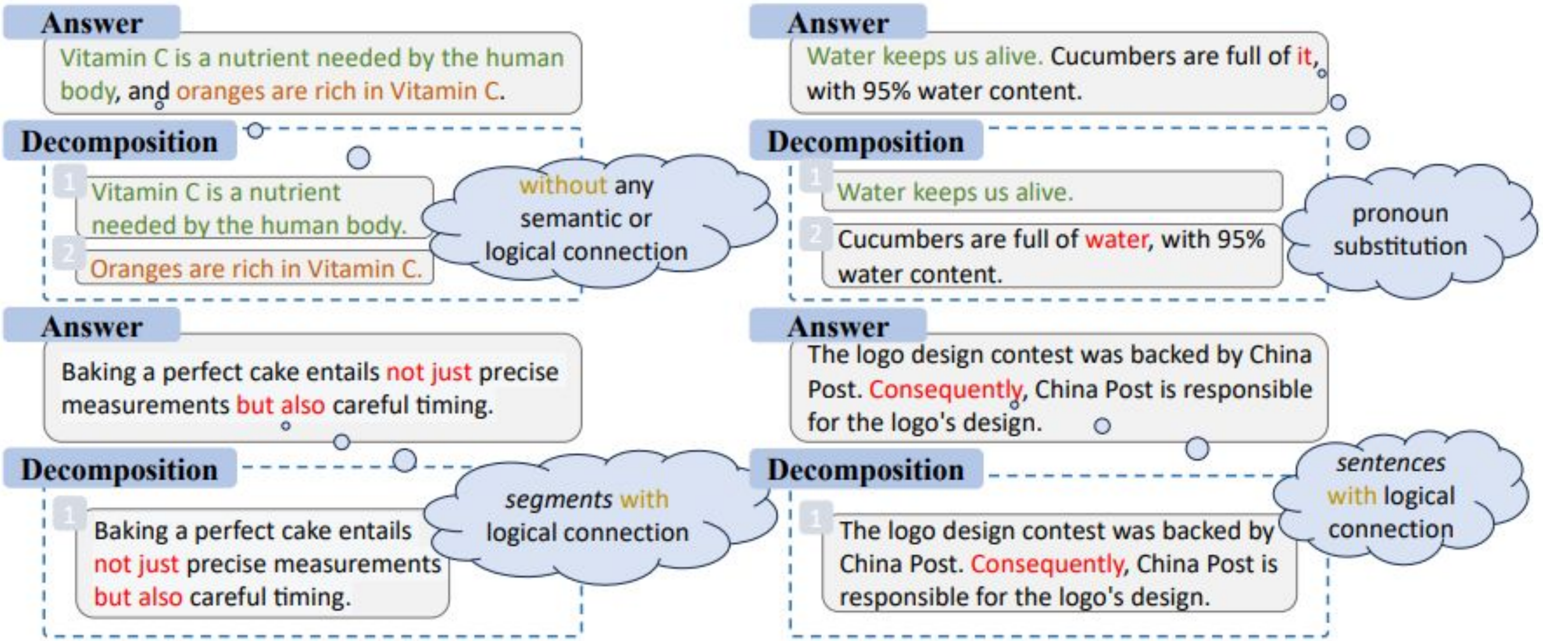
Methodology - Logic-Enhanced Factual Consistency Evaluation

- **Introduction**
  - Logical fallacy accounts for a considerable proportion of factual errors in real-world RAG scenarios.
  - Existing FCE pipelines neglect the logical connections between segments in the original answer, which may result in wrong factual consistency evaluation result for samples with logical fallacy.

Methodology - Logic-Enhanced Factual Consistency Evaluation

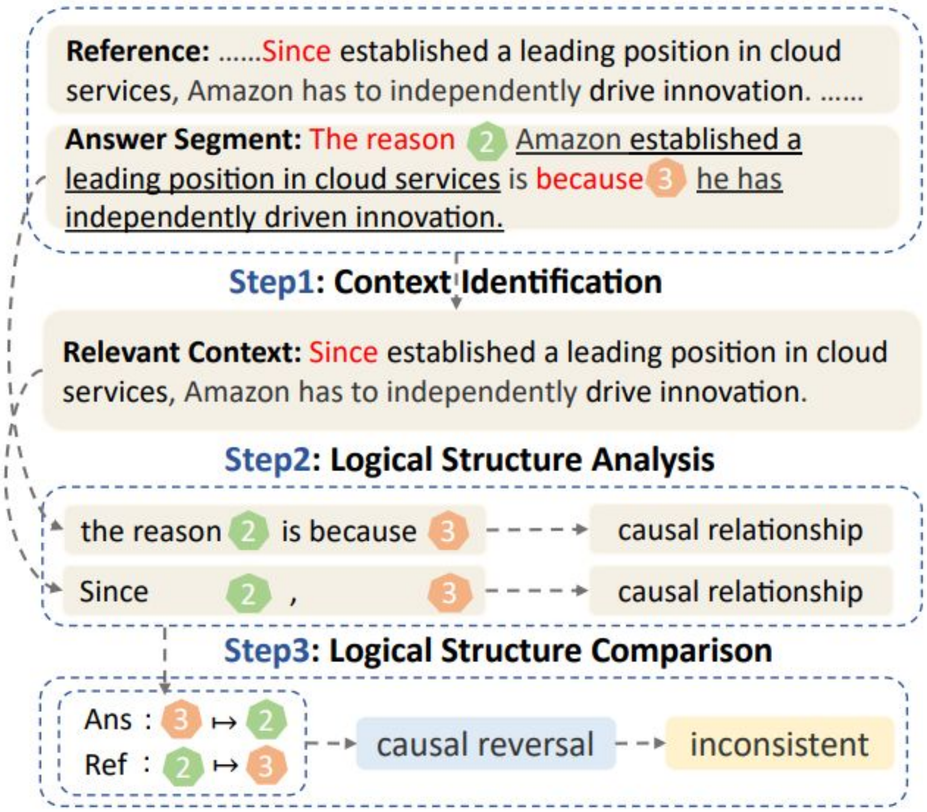
- **Logic-Preserving Answer Decomposition**
    - Logical Connection
    - Pronoun Substitution
    - Unique Format

Methodology - Logic-Enhanced Factual Consistency Evaluation



Methodology - Logic-Enhanced Factual Consistency Evaluation

- **Fact-Logic FCE**
  - Fact Consistency Evaluation
    - Informational Points Extraction
    - Context Identification
    - Fact Consistency Check
  - Logic Consistency Evaluation
    - Context Identification
    - Logical Structure Analysis
    - Logical Structure Comparison



Outline

- Introduction
- Related Works
- Methodology
  - Face4RAG Benchmark
  - Logic-Enhanced Factual Consistency Evaluation
- Experiment
- Conclusion
- References

Experiment

- **Baselines:**
    - FACTSCORE [31]
    - FELM [9]
    - Ragas [13]
    - Refchecker [18]

Experiment

- **Performance Comparison on Face4RAG - Synthetic Dataset**
  - L-Face4RAG shows significant performance improvement over other baselines
  - Particularly strong in detecting logical fallacy errors

Method	Total	Pos.	Negative samples									
			Hallu.	KCont.	KInve.	KConf.	KConc.	LOver.	LCaus.	LConf.	LIncl.	LOthe.
FACTSCORE(GPT-3.5)	70.36	37.31	90.45	**100**	94.44	55.56	94.29	78.57	64.29	68.00	46.34	86.07
FACTSCORE(GPT-4)	71.82	33.50	93.97	**100**	96.30	68.25	97.14	87.5	60.71	72.00	51.22	88.37
FELM	68.05	77.67	42.21	99.27	91.98	22.22	88.57	69.64	42.86	54.00	4.88	32.56
RAGAS(GPT-3.5)	69.59	70.81	76.89	98.54	71.60	49.21	87.14	54.46	39.29	48.00	34.15	44.19
RAGAS(GPT-4)	76.37	73.60	93.97	99.27	79.01	52.38	90.00	58.93	50.00	50.00	**53.66**	72.09
RefChecker	78.52	76.14	95.48	**100**	87.65	63.49	92.86	55.36	50.00	52.00	36.59	67.44
L-Face4RAG (Ours)	**93.38**	**96.19**	**96.98**	**100**	**98.77**	**76.19**	**98.57**	**90.18**	**92.86**	**80.00**	51.22	**90.70**

Experiment

- **Performance Comparison on Face4RAG - Real-world Dataset**
  - L-Face4RAG significantly outperforms other baseline FCE methods in overall accuracy

Method	Total	Baichuan2	ChatGLM3	GPT-3.5	GPT-4	Alpaca2 (CH)	Qwen
FACTSCORE(GPT-3.5)	53.33	54.0	55.5	47.5	51.5	59.0	52.5
FACTSCORE(GPT-4)	54.67	55.0	59.5	46.5	52.5	63.0	51.5
FELM	55.00	49.6	56.0	56.8	52.0	55.6	60.0
RAGAS(GPT-3.5)	65.92	64.5	68.5	64.5	60.0	65.0	73.0
RAGAS(GPT-4)	72.92	72.5	74.0	71.5	68.5	76.5	74.5
RefChecker	68.25	62.0	72.0	66.5	63.0	74.5	71.5
L-Face4RAG (Ours)	**87.75**	**90.0**	**88.0**	**81.5**	**86.0**	**93.5**	**87.5**

Experiment

- **Performance Comparison on Existing FCE Benchmark**
  - L-Face4RAG achieved SOTA (state-of-the-art) results on 6 out of 7 datasets

Method	Avg.	RAG		Summ.		Dial.		Fact Verif.
		RAGAS[13]	RefChecker[18]	FRANK[34]	SummEval[14]	Q^2[17]	DialFact[16]	VitaminC[37]
FACTSCORE(GPT-4)	70.5	70	61	80	65	74	72	71
FELM	74.2	71	63	70	<u>82</u>	<u>83</u>	**79**	72
RAGAS(GPT-4)	76.9	<u>88</u>	69	**87**	80	77	69	69
RefChecker	<u>78.4</u>	86	**73**	85	80	80	72	<u>73</u>
L-Face4RAG (Ours)	**84.2**	**91**	**73**	**87**	**90**	**84**	<u>77</u>	**88**

Experiment

- **Ablation Study**
  - Evaluating the Answer Decomposition Module. (A.D.)
  - Evaluating the Introduction of COT.(w/o COT)
  - Evaluating the Stage of Logical Consistency Evaluation. (w/o logi. eval)

	L-Face4RAG	A.D.	w/o COT	w/o logi.eval
Overall	93.38	76.44	79.60	88.99
-Positive	96.19	91.62	51.27	97.46
-Negative	92.15	69.83	91.93	85.30

Experiment

- **Ablation Study**
  - Removing the logical validation phase caused a **significant drop in logical performance**

		L-Face4RAG	A.D.	w/o logi. eval
Hallucination	Hallu.	96.98	90.45	96.98
Knowledge	KCont.	100.00	100.00	100.00
	KInve.	98.77	74.07	97.53
	KConf.	76.19	41.27	66.67
	KConc.	98.57	90.00	94.29
Logical	LOver.	90.18	42.86	83.93
	LCaus.	92.86	32.14	35.71
	LConf.	80.00	34.00	64.00
	LIncl.	51.22	31.71	29.27
	LOth.	90.70	44.19	65.12

Outline

Conclusion

- **Contribution**
  - Construct a comprehensive benchmark to enable the evaluation of FCE methods independent of the underlying LLM, which is called Face4RAG.
  - Propose a new FCE method called L-Face4RAG to better detect the logic fallacy in the examined answer

Outline

- Introduction
- Related Works
- Methodology
  - Face4RAG Benchmark
  - Logic-Enhanced Factual Consistency Evaluation
- Experiment
- Conclusion
- References

References

[1] 2019. https://gaokao.neea.edu.cn/xhtml1/report/19012/5987-1.htm. [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023). [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv preprint arXiv:2309.16609 (2023). [4] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. arXiv preprint arXiv:2306.04181 (2023). [5] Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. arXiv preprint arXiv:2309.10305 (2023). https://arxiv.org/abs/2309.10305 [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901. [7] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In Proceedings of the 2nd workshop on machine reading for question answering. 119–124. [8] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017). [9] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. arXiv preprint arXiv:2310.00741 (2023).

References

[10] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and MultiDomain Scenarios. arXiv preprint arXiv:2307.13528 (2023). [11] Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. arXiv preprint arXiv:2304.08177 (2023). https://arxiv.org/abs/2304.08177 [12] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In Proceedings of acl-08: Hlt. 1039–1047. [13] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217 (2023). [14] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics 9 (2021), 391–409. [15] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554 (2023). [16] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. DialFact: A benchmark for fact-checking in dialogue. arXiv preprint arXiv:2110.08222 (2021). [17] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Qˆ{2}: Evaluating Factual Consistency in KnowledgeGrounded Dialogues via Question Generation and Question Answering. arXiv preprint arXiv:2104.08202 (2021). [18] Xiangkun Hu, Dongyu Ru, Qipeng Guo, Lin Qiu, and Zheng Zhang. 2023. RefChecker for Fine-grained Hallucination Detection. (2023). https://github.com/amazon-science/RefChecker [19] Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282 (2020). [20] Ray S Jackendoff. 1992. Semantic structures. Vol. 18. MIT press. [21] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. Comput. Surveys 55, 12 (2023), 1–38.

References

[22] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, ZhihengLyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logicalfallacy detection. arXiv preprint arXiv:2202.13758 (2022).[23] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice:Real-world entailment for claims in wikipedia. arXiv preprint arXiv:2303.01432(2023).[24] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019.Evaluating the factual consistency of abstractive text summarization. arXivpreprint arXiv:1910.12840 (2019).[25] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022.SummaC: Re-visiting NLI-based models for inconsistency detection in summa-rization. Transactions of the Association for Computational Linguistics 10 (2022),163–177.[26] Barrett Martin Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. 2023. Fastand Accurate Factual Inconsistency Detection Over Long Documents. arXivpreprint arXiv:2310.13189 (2023).[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, VladimirKarpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, TimRocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020),9459–9474.[28] Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approachesin natural language processing: A survey. Ai Open 3 (2022), 71–90.[29] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability ingenerative search engines. arXiv preprint arXiv:2304.09848 (2023).[30] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020.On faithfulness and factuality in abstractive summarization. arXiv preprintarXiv:2005.00661 (2020).[31] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang WeiKoh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore:Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Gener-ation. arXiv preprint arXiv:2305.14251 (2023).[32] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov,Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023.Generating benchmarks for factuality evaluation of language models. arXivpreprint arXiv:2307.06908 (2023).

References

[33] OpenAI. 2022. Chatgpt blog post. https://openai.com/blog/chatgpt.[34] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Under-standing factuality in abstractive summarization with FRANK: A benchmark forfactuality metrics. arXiv preprint arXiv:2104.13346 (2021).[35] Domina Petric. 2020. Logical Fallacies. On-line Article (preprint), doi 10 (2020).[36] Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, YankaiLin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. WebCPM: Interac-tive Web Search for Chinese Long-form Question Answering. arXiv preprintarXiv:2305.06849 (2023). [37] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! ro-bust fact verification with contrastive evidence. arXiv preprint arXiv:2103.08541(2021).[38] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, StephenRoller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blender-bot 3: a deployed conversational agent that continually learns to responsiblyengage. arXiv preprint arXiv:2208.03188 (2022).[39] Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu,Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022.Understanding factual errors in summarization: Errors, summarizers, datasets,error detectors. arXiv preprint arXiv:2205.12854 (2022).[40] Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evalu-ating accuracy in data-to-text systems. arXiv preprint arXiv:2011.03992 (2020).[41] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kul-shreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al.2022. Lamda: Language models for dialog applications. arXiv preprintarXiv:2201.08239 (2022). [42] Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and YueZhang. 2023. Evaluating open question answering evaluation. arXiv preprintarXiv:2305.12421 (2023).[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi,Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reason-ing in large language models. Advances in Neural Information Processing Systems35 (2022), 24824–24837.

References

[44] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding,Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An openbilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022).[45] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Eval-uating Factual Consistency with a Unified Alignment Function. arXiv preprintarXiv:2305.16739 (2023).[46] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and YoavArtzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprintarXiv:1904.09675 (2019).[47] Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and Detect-ing Fine-grained Factual Errors for Dialogue Summarization. arXiv preprintarXiv:2305.16548 (2023).

SoWork – Lottery Post Recognition

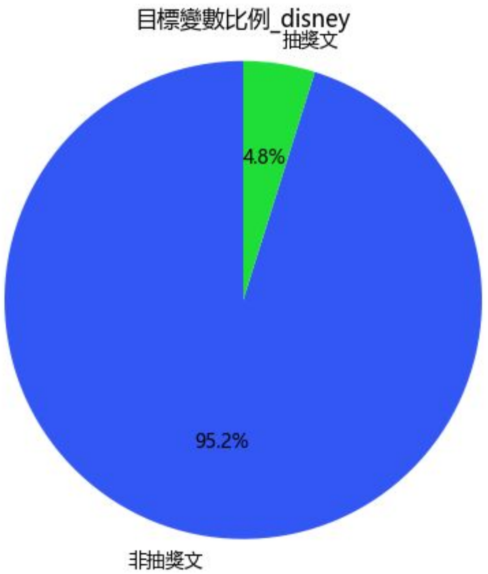
Presenter: Cheng Jhe Lee
National Cheng Kung University

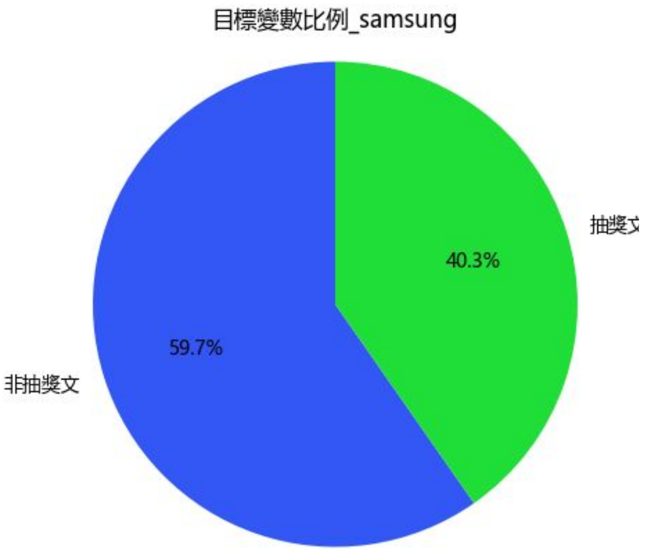
Data Analysis

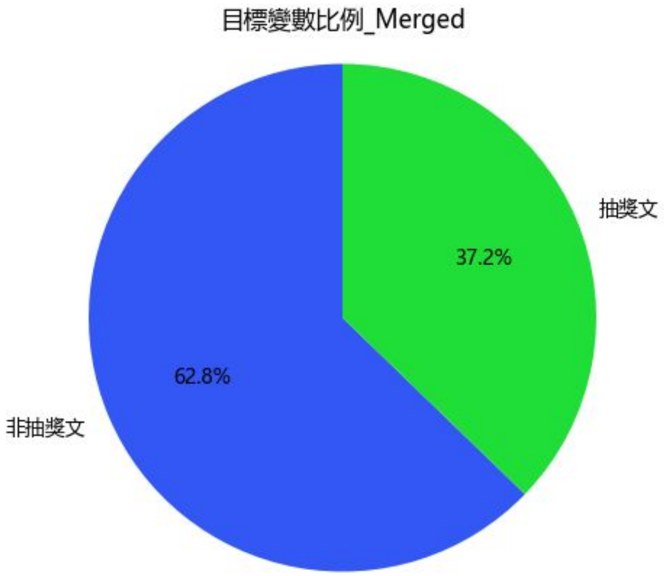
- **Data Information**
  - Two dataset, Disney+ dataset and Samsung dataset
  - Disney+ has **30 columns**, while samsung dataset only has **25 columns**, the former has 5 more columns, `['C', 'URL', '合併', '日期', '時間']`
  - Merged dataset has **25 columns** with **32809 rows**

Data Analysis

- **Label Distribution**







Data Analysis

- **Feature Extraction**

Most significant	**討論串總則數(OpView收錄回文文章數)**
Important	來源、作者、頻道屬性標籤、頻道內容標籤、網站、按讚/觀看、正面強度、負面強度
Minor	分享/轉貼數/評級、FB_讚、FB_大心、FB_哈、FB_哇、FB_嗚、FB_怒、中立強度、監測主題、發布時間、標題、內容、主文/回文、情緒標記、原始連結

Conclusion

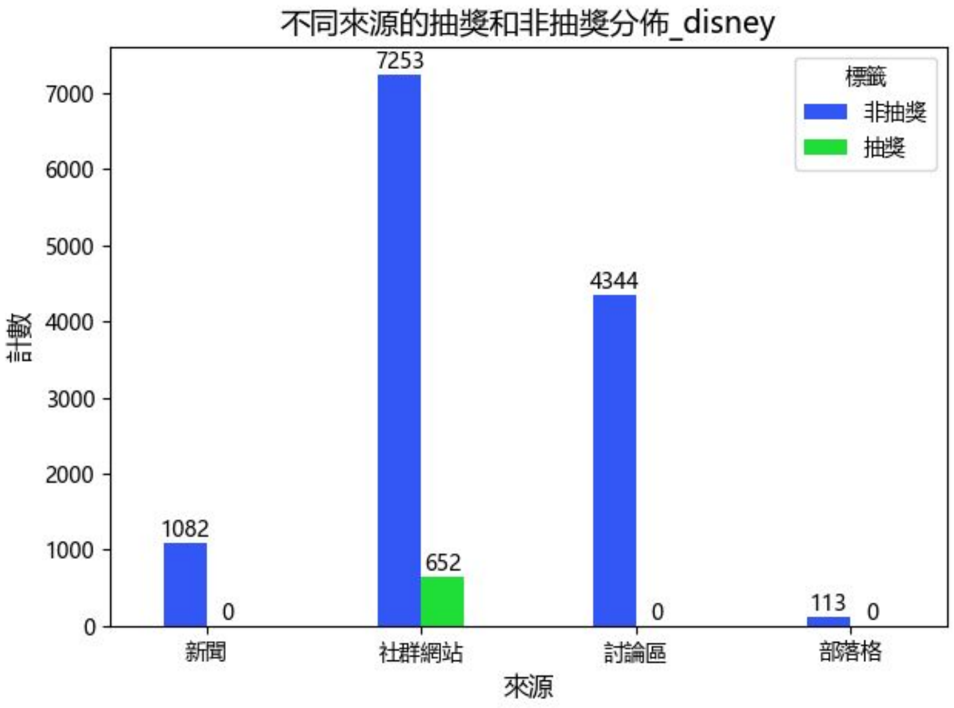
- Merged dataset has **25 columns** with **32809 rows**

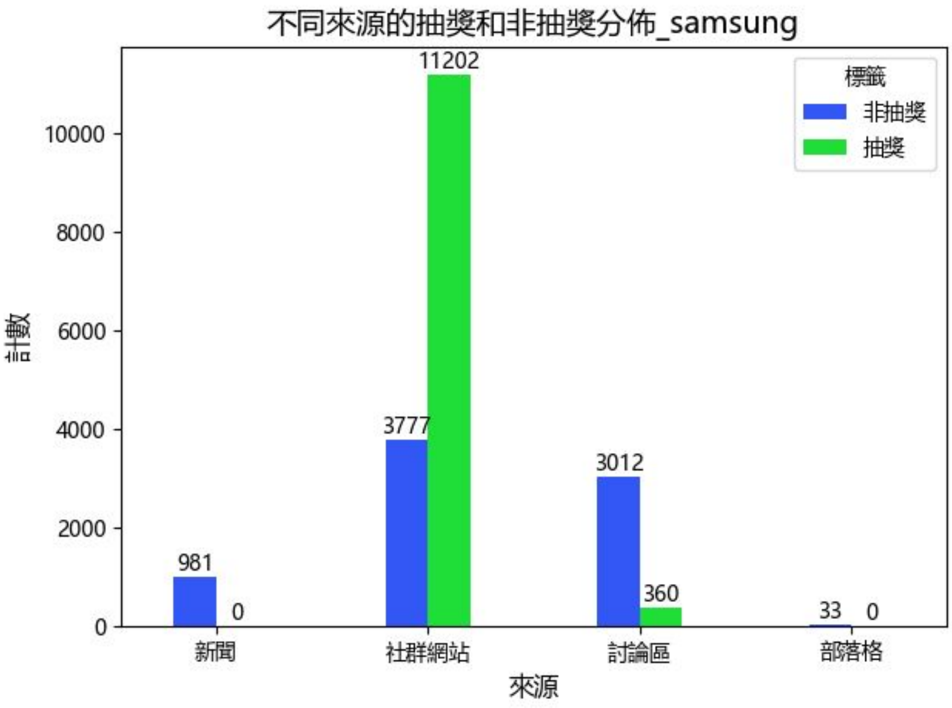
- Most significant feature is '**討論串總則數(OpView收錄回文文章數)**', which is **consistent with the description provided by Sowork**

- After classifying three datasets with a decision tree and analyzing their precision, accuracy, and recall, the lowest score was **93%**

Appendix

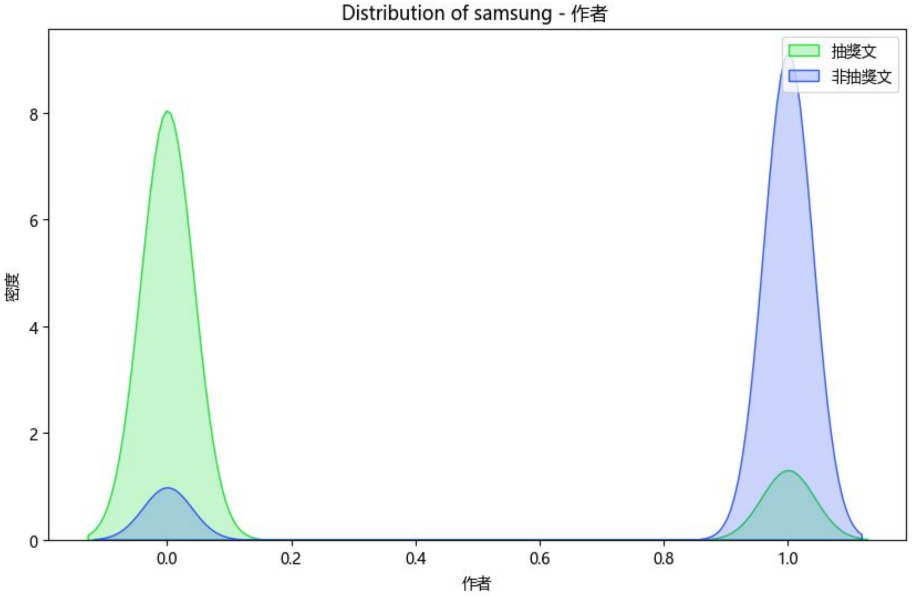
- **column '來源'**

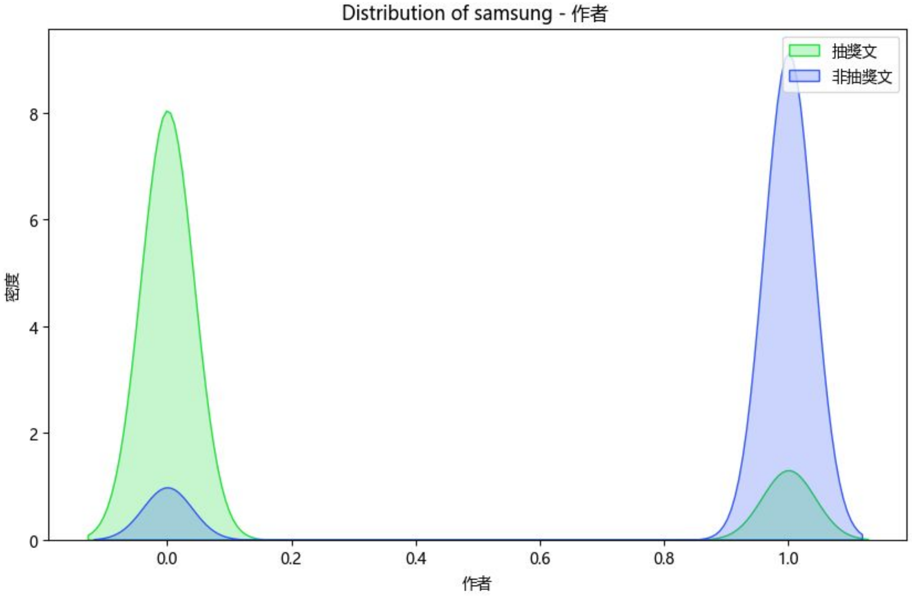




Appendix

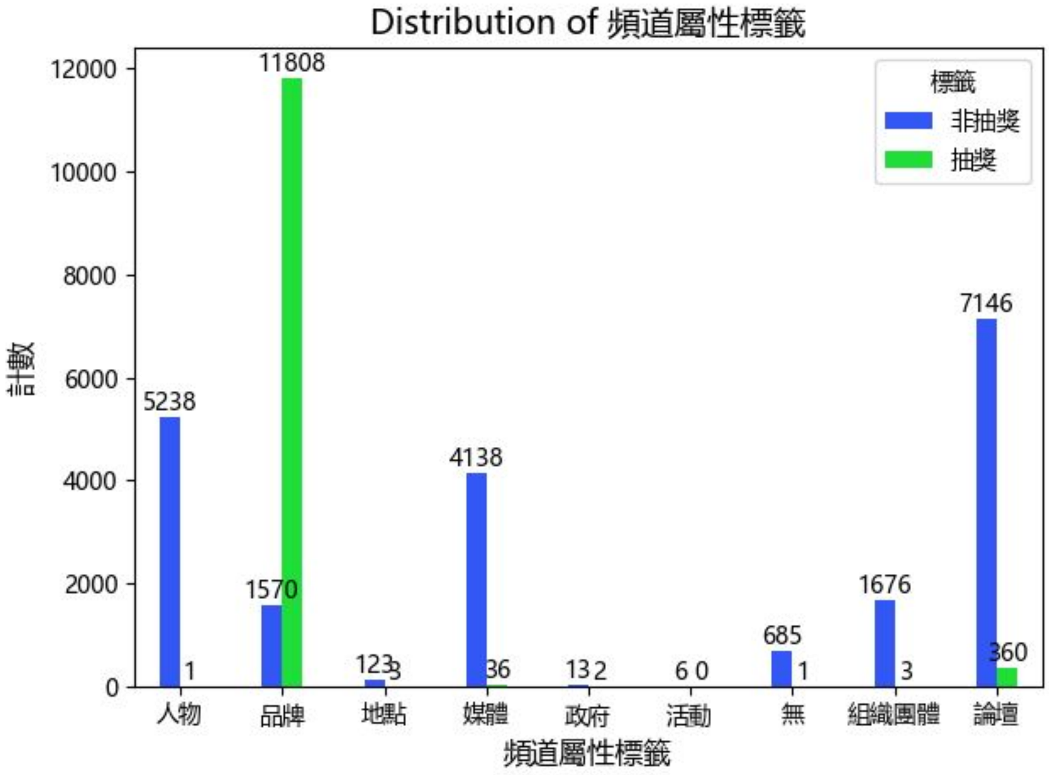
- column '作者'





Appendix

- **(merged) column '頻道屬性標籤'**



Appendix

- **Decision tree visualization**