Imitating Human Reasoning to Extract 5W1H in News

Carlos Muñoz
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago, Chile
carlos.munoz@uc.cl

Marcelo Mendoza*
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago, Chile
marcelo.mendoza@uc.cl

Hans Lobel
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago, Chile
halobel@uc.cl

Brian Keith
Department of Computing and Systems Engineering
Universidad Católica del Norte
Antofagasta, Chile
brian.keith@ucn.cl

Abstract

Extracting key information from news articles is crucial for advancing search systems. Historically, the 5W1H framework, which organises information based on 'Who', 'What', 'When', 'Where', 'Why', and 'How', has been a predominant method in digital journalism empowering search tools. The rise of Large Language Models (LLMs) has sparked new research into their potential for performing such information extraction tasks effectively. Our study examines a novel approach to employing LLMs in the 5W1H extraction process, particularly focusing on their capacity to mimic human reasoning. We introduce two innovative Chain-of-Thought (COT) prompting techniques to extract 5W1H in news: extractive reasoning and question-level reasoning. The former directs the LLM to pinpoint and highlight essential details from texts, while the latter encourages the model to emulate human-like reasoning at the question-response level. Our research methodology includes experiments with leading LLMs using prompting strategies to ascertain the most effective approach. The results indicate that COT prompting significantly outperforms other methods. In addition, we show that the effectiveness of LLMs in such tasks depends greatly on the nature of the questions posed.

CCS Concepts

• **Computing methodologies → Information extraction**.

Keywords

5W1H, LLM, imitative reasoning, news

ACM Reference Format:
Carlos Muñoz, Marcelo Mendoza, Hans Lobel, and Brian Keith. 2025. Imitating Human Reasoning to Extract 5W1H in News. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715532

1 Introduction

The 5W1H framework is a fundamental approach for analyzing and organizing information in the news landscape. It is commonly employed in journalism [7], research [3], and problem-solving [1] to ensure that all critical aspects of a situation or event are comprehensively covered [2]. In the context of search, applying 5W1H allows systems to better identify and classify relevant news articles by focusing on these key elements [6]. Each component serves a role in enhancing the granularity of the information retrieved [8].

COT prompting has become increasingly significant for reasoning tasks [11]. COT refers to a process in which the model generates intermediate steps that guide it toward a logical conclusion [10]. This step-by-step reasoning approach enhances the model's ability to tackle complex problems that require multiple stages of inference. By leveraging COT in LLMs, it is possible to break down intricate queries into manageable parts, which can result in more accurate and reliable outcomes [9].

We introduce COT prompts for 5W1H extraction in news. We conducted an experimental framework to evaluate different prompting strategies on some of the most powerful LLMs currently available. Using a news dataset annotated with text spans, we measured the consistency of the LLMs' responses through standard text-matching metrics.

To our knowledge, **this is the first study to compare COT and few shot prompts for 5W1H extraction in English news**. Our principal contributions are outlined as follows:

- We introduce two COT strategies inspired by human reasoning specifically designed to extract 5W1H in news.
- Our experimental findings demonstrate that the strategies collectively yield satisfactory outcomes for this task, with certain strategies and models proving to be more effective for specific questions.

2 Methodology

2.1 Task definition

The extraction of 5W1H is the identification of relevant units of information from a text based on six key questions that help identify

[*] Corresponding author.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715532

essential components of an informative text. The questions that form this framework are:

- *What*, which describes the key facts, circumstances, and/or actions mentioned in the news;
- *Who*, which identifies the main subject or entity involved in the news;
- *When*, specifying the relevant time and/or date when the events occurred;
- *Where*, which indicates the location or place where the events took place;
- *Why*, explaining the reasons or causes behind the event; and
- *How*, which outlines the manner or method in which the event unfolded.

In the context of news reporting, the descriptive text of a news story typically includes a headline, a lead, and the body of the news. It is understood that both the headline and lead describe the **event**, while the body of the news provides a detailed description of the event details. The headline summarises the essence of the news, and the lead gives additional context, whereas the body delves into the event's particulars.

Automatic 5W1H extraction involves identifying text spans that answer the six questions described above.

2.2 The reasoning behind 5W1H extraction

The first COT strategy, described in Fig. 1, was developed to emulate reasoning based on the principle of information extraction. To implement this, the LLM is provided with guidelines that instruct when to remove a sentence and when to retain it for further analysis. The inclusion/exclusion guidelines are:

- Remove sentences related to additional content, background information, or descriptive elements within the 'NEWS_BODY' that do not directly relate to the main 'EVENT', and replace such sentences with '[...]'.
- Ensure that the resulting 'NEWS_BODY' remains focused on addressing the core 5W1H questions about the main 'EVENT'. Give priority to content that clearly answers these questions.
- Preserve the original wording of the article as much as possible, with the only modification being the replacement of irrelevant sentences.
- Finally, if no irrelevant sentences are identified or if the analysis proves challenging due to the provided content, leave the 'NEWS_BODY' unchanged.

The second COT strategy used in this study, described in Fig. 2, incorporated specific guidelines tailored to each of the 5W1H questions. These guidelines follow the instructions used to train human annotators within the 5W1H framework, according to [4]. We embedded the rationale for applying these guidelines in the examples provided to the LLM. This approach allows the LLM to link reasoning to a finer level of granularity, depending on the question it is addressing.

3 Experiments

3.1 Design of experiments

This study addresses three research questions:

- Q1: Which LLM model is the most effective for 5W1H extraction?
- Q2: What prompting strategy is the most effective for 5W1H extraction?
- Q3: Are there more effective prompt strategies for certain 5W1H questions?

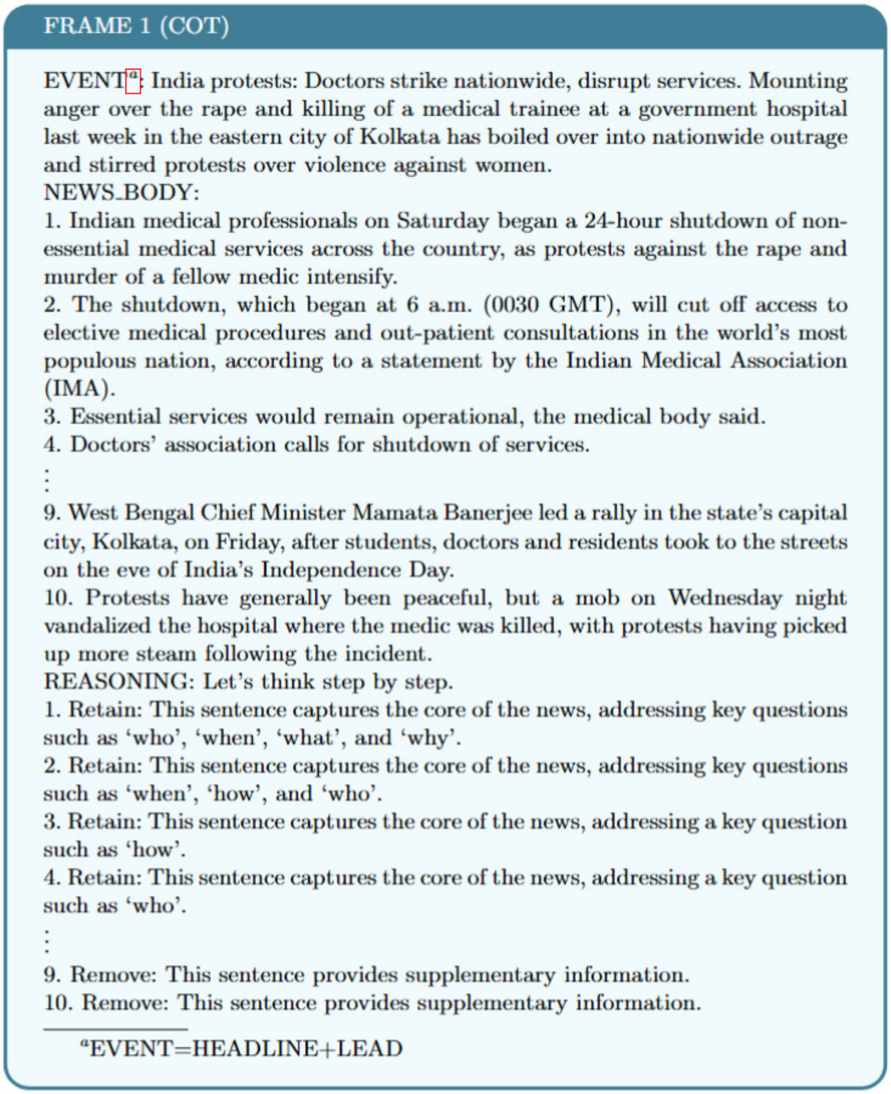
To address these questions, we evaluated prompting strategies based on the responses provided by the most prominent LLMs available. We assessed GPT-4o (OpenAI, version release August 6, 2024, https://platform.openai.com/docs/models/gpt-4o), Claude-3-5-sonnet (Anthropic, version release June 20, 2024, https://claude.ai/), and Gemini-1.5-Pro (Google DeepMind, version release May 24, 2024, https://deepmind.google/technologies/gemini/pro/).

3.2 Prompt engineering

We evaluate three prompting strategies to extract 5W1H in news: COT, zero-shot and few-shot prompting. For each of these strategies, we used prompts that defined the task to be solved and incorporated elements that enriched the task's contextual description. The prompts were designed incrementally, each new prompt adding more information to complete the task description, providing additional instructions, and including examples. The set of prompts used in this study is:

- P1 (zero-shot) is a prompt that defines the extraction task. In the system role, general instructions are given, general rules are established, and it is emphasized that the responses must be directly based on the provided content. In the user role, the news article is provided, including the headline, lead, and body.
- P2 (zero-shot) builds upon P1 by adding a detailed description of each 5W1H element.
- P3 (zero-shot) extends P2 by adding the instruction, "Only use excerpts from the provided context.
- P4 (one-shot) adds to P3 an example that includes a news article and the expected responses.
- P5 (few-shot) builds on P4 by adding a second example along with the expected answers.
- P6 (few-shot) builds on P5 by adding a third example with the corresponding expected answers.
- P7 (Extractive COT, ours) defines guidelines for removing irrelevant text. It specifically asks to remove sentences in the body of the text that do not directly relate to the headline and lead of the news. It includes one example. After irrelevant text is filtered out, the 5W1H extraction is performed using one-shot prompting.
- P8 (Extractive COT, ours) mirrors P7 but uses few-shot prompting based on two examples.
- P9 (Extractive COT, ours) follows the same logic as P8 but incorporates three examples.
- P10 (Question-level COT, ours) introduces complex reasoning for each question using one example. The reasoning per question is based on annotation guidelines used in [4].
- P11 (Question-level COT, ours) mirrors P10 but with two examples, making it COT few-shot.

To construct the few-shot prompts, five examples not included in the dataset but studied in [4] were used.



FRAME 1 (COT)

EVENT[a]: India protests: Doctors strike nationwide, disrupt services. Mounting anger over the rape and killing of a medical trainee at a government hospital last week in the eastern city of Kolkata has boiled over into nationwide outrage and stirred protests over violence against women.
NEWS_BODY:
1. Indian medical professionals on Saturday began a 24-hour shutdown of non-essential medical services across the country, as protests against the rape and murder of a fellow medic intensify.
2. The shutdown, which began at 6 a.m. (0030 GMT), will cut off access to elective medical procedures and out-patient consultations in the world's most populous nation, according to a statement by the Indian Medical Association (IMA).
3. Essential services would remain operational, the medical body said.
4. Doctors' association calls for shutdown of services.
⋮
9. West Bengal Chief Minister Mamata Banerjee led a rally in the state's capital city, Kolkata, on Friday, after students, doctors and residents took to the streets on the eve of India's Independence Day.
10. Protests have generally been peaceful, but a mob on Wednesday night vandalized the hospital where the medic was killed, with protests having picked up more steam following the incident.
REASONING: Let's think step by step.
1. Retain: This sentence captures the core of the news, addressing key questions such as 'who', 'when', 'what', and 'why'.
2. Retain: This sentence captures the core of the news, addressing key questions such as 'when', 'how', and 'who'.
3. Retain: This sentence captures the core of the news, addressing a key question such as 'how'.
4. Retain: This sentence captures the core of the news, addressing a key question such as 'who'.
⋮
9. Remove: This sentence provides supplementary information.
10. Remove: This sentence provides supplementary information.

[a] EVENT=HEADLINE+LEAD

Figure 1: Extractive reasoning COT prompt used for 5W1H extraction.

To favor reproducibility the prompts are available in: https://github.com/cmunhozc/5W1H-prompt-strategies.

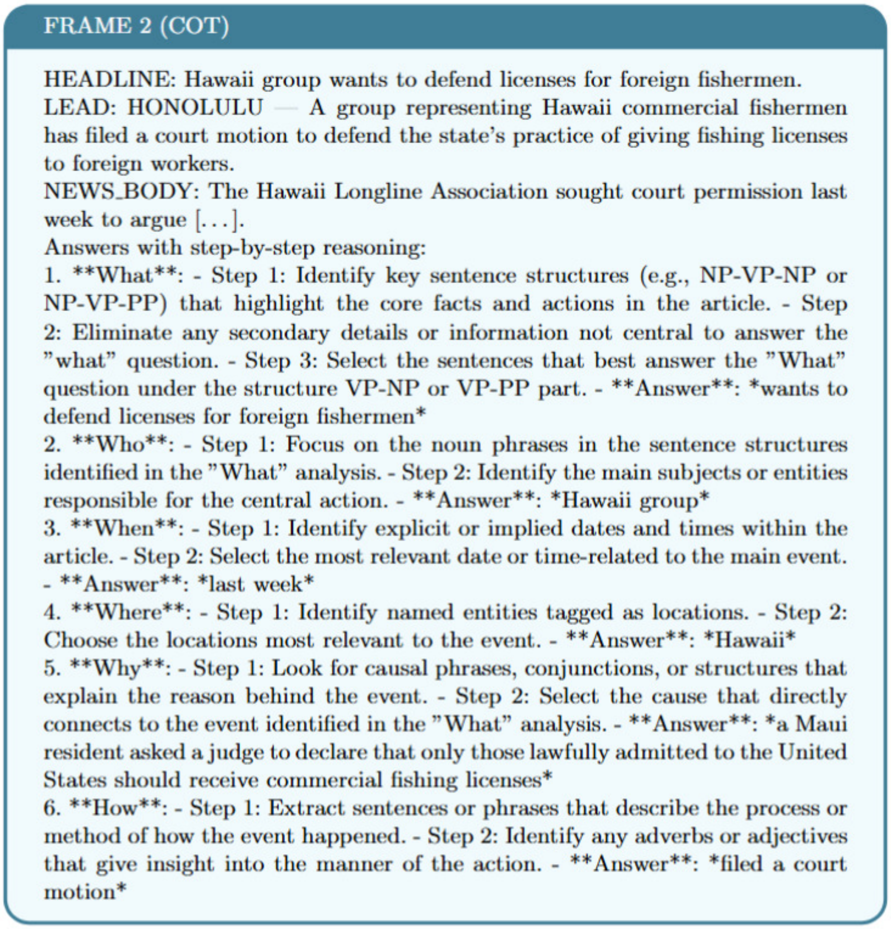
3.3 Dataset

We explore our research questions using the Giveme5W1H dataset [2], which gathers news across various topics, including politics, business, international affairs, and sports. The news articles come from various web sources and include various publishers, such as the Daily Mail, The Sun, The Independent, Mirror, BBC, Telegraph, and others. For each article, the dataset contains key elements such as the 'headline', 'lead', and the 'news body'. Each news article is annotated according to the 5W1H framework, forming the gold standard partition of the dataset, which comprises a total of 96 news articles (Data available at: https://github.com/fhamborg/Giveme5W1H/tree/master/Giveme5W1H/examples/datasets/gold_standard/data).

3.4 Results

To address Q1, we assessed the correspondence between the responses generated by the LLMs and Giveme5W1H. We used ROUGE [5], including ROUGE-1, which measures unigram overlap, ROUGE-2, which captures bigram overlap, and ROUGE-L which focus on the longest common subsequence. Furthermore, we included F_BERT [12], which evaluates similarity using contextual embeddings. Table 1 shows these results.

Upon reviewing the best results per metric for each LLM (bold fonts), we observe that COT prompts achieve better outcomes. Regarding each metric, while ROUGE-based metrics demonstrate how closely outcomes align with reference texts, the F_BERT metric reveals that semantic alignment with reference responses is also achieved. Table 1 shows that both Gemini-1.5 and GPT-4o improved ROUGE results using COT Frame 2 (P10-P11) and better semantic



FRAME 2 (COT)

HEADLINE: Hawaii group wants to defend licenses for foreign fishermen.
LEAD: HONOLULU — A group representing Hawaii commercial fishermen has filed a court motion to defend the state's practice of giving fishing licenses to foreign workers.
NEWS_BODY: The Hawaii Longline Association sought court permission last week to argue [. . .].
Answers with step-by-step reasoning:
1. **What**: - Step 1: Identify key sentence structures (e.g., NP-VP-NP or NP-VP-PP) that highlight the core facts and actions in the article. - Step 2: Eliminate any secondary details or information not central to answer the ”what” question. - Step 3: Select the sentences that best answer the ”What” question under the structure VP-NP or VP-PP part. - **Answer**: *wants to defend licenses for foreign fishermen*
2. **Who**: - Step 1: Focus on the noun phrases in the sentence structures identified in the ”What” analysis. - Step 2: Identify the main subjects or entities responsible for the central action. - **Answer**: *Hawaii group*
3. **When**: - Step 1: Identify explicit or implied dates and times within the article. - Step 2: Select the most relevant date or time-related to the main event. - **Answer**: *last week*
4. **Where**: - Step 1: Identify named entities tagged as locations. - Step 2: Choose the locations most relevant to the event. - **Answer**: *Hawaii*
5. **Why**: - Step 1: Look for causal phrases, conjunctions, or structures that explain the reason behind the event. - Step 2: Select the cause that directly connects to the event identified in the ”What” analysis. - **Answer**: *a Maui resident asked a judge to declare that only those lawfully admitted to the United States should receive commercial fishing licenses*
6. **How**: - Step 1: Extract sentences or phrases that describe the process or method of how the event happened. - Step 2: Identify any adverbs or adjectives that give insight into the manner of the action. - **Answer**: *filed a court motion*

Figure 2: Question-level reasoning COT prompt used for 5W1H extraction.

matches using COT Frame 1 (P7-P9). Claude-3.5 exhibits a different behavior, enhancing both lexical and semantic matching when employing question-level reasoning (COT Frame 2).

Regarding the global results and concerning **Q1**, **the best performances are obtained by GPT-4o and Claude-3.5**. Concerning **Q2**, **the evidence indicates a strong trend that favors prompts that incorporate reasoning**. These prompts significantly outperform zero (P1-P3), one (P4), or few-shot learning prompts (P5-P6). **We conclude that no single prompt guarantees the best overall results but COT prompts surpasses the quality of the results obtained using other techniques.**

To address **Q3**, we use the same metrics but this time disaggregated by question type. We present the results achieved by each LLM according to the prompts that yielded their best performance. These results are shown in Table 2.

Table 2 shows that, generally, **the question where LLMs perform best is 'who'**. The only model that deviates from this pattern is GPT-4o, which performs better in the 'where' question. Table 2 also shows that **the most challenging question type is 'how'**. For the 'who' question, Claude-3.5 and Gemini-1.5 achieve their best results using question-level reasoning (P10-P11), while GPT-4o do so using extractive reasoning (P8-P9). Globally, the experiments demonstrate that no single model achieves the best results across all six questions. On the one hand, GPT-4o effectively answers the 'when', 'where', and 'why' questions. In the first two, the best strategy relies on question-level reasoning (P10). However, for the 'why' question, the best approach used by GPT-4o is based on extractive reasoning (P7). On the other hand, for the 'what', 'who', and 'how' questions, the best-performing model is Claude-3.5. In this case, the most effective strategy consistently relies on question-level reasoning (P11). Consequently, regarding Q3, the evidence shows that question-level reasoning (P10-P11) is superior for five of the six questions in the framework. In contrast, extractive reasoning (P7) is more effective for the remaining 'why' question.

4 Conclusion

Mimicking human reasoning using COT prompting is the most effective strategy for 5W1H extraction. We demonstrate that GPT-4o and Claude 3.5 achieve the best results in the task. We also observe that the effectiveness of the prompts depends on both the specific LLM used and the type of question being asked. Furthermore, the

Table 1: Evaluation of prompting techniques for 5W1H extraction from Giveme5W1H data using LLMs. Bold fonts indicate the best results per metric for each LLM. The best global results per metric are depicted in red.

	P#	ROUGE-2	ROUGE-L	ROUGE-1	F_BERT
GPT-4o	P1	0.166	0.293	0.305	0.866
	P2	0.193	0.297	0.309	0.868
	P3	0.219	0.367	0.377	0.878
	P4	0.196	0.331	0.346	0.870
	P5	0.206	0.346	0.360	0.872
	P6	0.220	0.376	0.389	0.877
	P7	0.244	0.414	0.421	**0.892**
	P8	0.249	0.421	0.428	**0.892**
	P9	0.243	0.410	0.417	0.890
	P10	**0.260**	**0.427**	**0.437**	0.888
	P11	0.232	0.382	0.392	0.881
Claude-3.5	P1	0.129	0.226	0.243	0.862
	P2	0.153	0.257	0.274	0.869
	P3	0.173	0.292	0.310	0.876
	P4	0.145	0.250	0.269	0.868
	P5	0.164	0.278	0.297	0.874
	P6	0.173	0.330	0.348	0.876
	P7	0.202	0.320	0.338	0.882
	P8	0.196	0.315	0.330	0.882
	P9	0.195	0.318	0.333	0.882
	P10	0.236	0.375	0.387	**0.895**
	P11	**0.248**	**0.412**	**0.425**	0.893
Gemini-1.5	P1	0.134	0.238	0.254	0.860
	P2	0.161	0.295	0.311	0.873
	P3	0.160	0.288	0.302	0.873
	P4	0.126	0.233	0.244	0.858
	P5	0.139	0.249	0.260	0.860
	P6	0.149	0.279	0.290	0.864
	P7	0.204	0.326	0.339	**0.882**
	P8	0.210	0.324	0.337	0.878
	P9	0.200	0.315	0.328	0.877
	P10	**0.226**	**0.351**	**0.365**	0.872
	P11	0.174	0.282	0.291	0.861

Table 2: Best results obtained by each LLM for every 5W1H question. We highlight in bold fonts the best result per metric for each LLM. The red color marks the overall best performance for each question.

	Query	P#	ROUGE-2	ROUGE-L	ROUGE-1	F_BERT
GPT-4o	what	P8	0.276	0.383	0.404	0.885
	who	P9	**0.499**	0.638	0.641	0.916
	when	P10	0.292	0.112	0.570	0.906
	where	P10	0.310	**0.680**	**0.690**	**0.930**
	why	P7	0.234	0.317	0.321	0.880
	how	P10	0.073	0.117	0.121	0.840
Claude-3.5	what	P11	0.280	0.361	0.382	0.886
	who	P11	**0.514**	**0.650**	**0.677**	**0.927**
	when	P11	0.233	0.558	0.561	0.913
	where	P11	0.246	0.497	0.517	0.919
	why	P11	0.131	0.280	0.285	0.874
	how	P11	0.085	0.125	0.130	0.843
Gemini-1.5	what	P10	0.188	0.285	0.305	0.866
	who	P10	**0.401**	**0.550**	**0.573**	**0.896**
	when	P10	0.236	0.416	0.431	0.873
	where	P10	0.265	0.462	0.474	0.886
	why	P10	0.175	0.247	0.255	0.864
	how	P10	0.094	0.147	0.153	0.845

study reveals that extractive reasoning is more suitable for addressing the 'why' question, whereas the other five questions are better approached using question-level reasoning.

Future work could focus on extending our proposal to other more complex prompting strategies and evaluate the potential applicability of these findings to other domains. such as event ontology population [8], systematic literature review [3] or the detection of real events from multimedia [6].

Acknowledgments

M. Mendoza was funded by the National Agency of Research and Development (ANID), grant FONDECYT 1241462. H. Lobel and M. Mendoza acknowledge the support of the National Center of Artificial Intelligence (ANID grant FB210017) and the Millennium Institute for Foundational Research on Data (ANID grant ICN17_002). B. Keith was funded by the UCN grant 202311010033-VRIDT-UCN.

References

[1] Turkay Dereli and Alptekin Durmusoglu. 2010. An integrated framework for new product development using who-when-where-why-what-how (5W1H), theory of inventive problem solving and patent information–a case study. *International Journal of Industrial and Systems Engineering* 5, 3 (2010), 354–365.

[2] Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. *arXiv preprint arXiv:1909.02766* (2019).

[3] Changjiang Jia and Yuen Tak Yu. 2013. Using the 5W+ 1H model in reporting systematic literature review: A case study on software testing for cloud computing. In *2013 13th International Conference on Quality Software*. IEEE, 222–229.

[4] Brian Keith, Michael Horning, and Tanushree Mitra. 2020. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. *Computational Journalism C+ J* (2020).

[5] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[6] Siraj Mohammed, Fekade Getahun, and Richard Chbeir. 2021. 5W1H Aware Framework for Representing and Detecting Real Events from Multimedia Digital Ecosystem. In *Advances in Databases and Information Systems. ADBIS 2021*. Springer, Cham, 78–90. https://doi.org/10.1007/978-3-030-82472-3_6

[7] Miki Tanikawa. 2017. What is news? What is the newspaper? The physical, functional, and stylistic transformation of print newspapers, 1988–2013. *International Journal of Communication* 11 (2017), 22.

[8] Wenbin Wang, Lei Zhang, Juanzi Li, and Yang Zhang. 2012. Chinese News Event 5W1H Semantic Elements Extraction for Event Ontology Population. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, 197–202. https://doi.org/10.1145/2187980.2187997

[9] Yu Wang, Shiwan Zhao, Zhihu Wang, Heyuan Huang, Ming Fan, Yubo Zhang, Zhixing Wang, Haijun Wang, and Ting Liu. 2024. Strategic Chain-of-Thought: Guiding Accurate Reasoning in LLMs through Strategy Elicitation. *arXiv preprint arXiv:2409.03271* (2024). https://arxiv.org/abs/2409.03271 arXiv:2409.03271.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Curran Associates, Inc., 1–12. https://arxiv.org/abs/2201.11903

[11] Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of Thought Prompting Elicits Knowledge Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 6519–6534. https://doi.org/10.18653/v1/2023.findings-acl.408

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html