DEMON

Detect and Evaluate Manipulation of ONline information

Deliverable number: **WP2**

Content-based analysis of stories and sub-narratives (UNICA)
M6–M18

Università degli Studi di Cagliari

Project Acronym:	DEMON
Project Title:	Detect and Evaluate Manipulation of ONline information
Project Type:	PRIN 2022
Project Code:	2022BAXSPY
Project Website:	`https://demon.unica.it/`
Workpackage:	**WP2:** Content-based analysis of stories and sub-narratives
Deliverable Type:	Report (R)
Dissemination Level:	Public (PU)
Authors:	Maurizio Atzori, Cecilia Di Ruberto, Andrea Loddo, Davide Mura, Manuela Sanguinetti, Marco Usai (UNICA)

Contents

1 Introduction to WP2

The following contains the contributions of the DEMON project falling into the area of Work Package 2. During this period, WP2 activities concentrated on advancing narrative modeling and content analysis through several key initiatives:

- Multimodal Fake News Detection (Section 2): We expanded the Themis architecture to handle multimodal data, validating it on the Fakeddit and ReCOVery datasets. By integrating LoRA and advanced data augmentation, the system achieved competitive, state-of-the-art performance, particularly on the ReCOVery benchmark.
- Comparison of Multimodal Models (Section 3): an in-depth study – initiated as parte of WP1 and then completed in this phase of the project – was carried out on multimodal binary fake news classification, comparing the Themis and HAMMER architectures on the DGM4, Fakeddit, and Multi Fake Detective (MFD) datasets. Extensions with ALBEF were also tested to improve cross-modal alignment. These activities were completed on schedule, providing empirical evidence useful for the next phase.
- Data-Centric Propaganda Analysis (Section 4): A comprehensive study of major propaganda detection datasets was conducted. By examining annotation schemes and linguistic variety, the research highlighted how data quality directly dictates model efficacy, leading to more robust standards for dataset construction.
- Sub-Narrative Classification (Section 5): A new multilingual system was engineered to classify sub-narratives using a hierarchical taxonomy. This pipeline utilizes BERT-based contextual embeddings and neural networks to aggregate sentence-level predictions into document-level insights.
- Entity Framing and Role Labeling (Section 6): Leveraging LLaMA 3 (8B) with QLoRA, we developed a system to identify entity roles (e.g., victim vs. perpetrator) across five languages. Focusing on high-stakes topics like the Ukraine war, the model reached 84% accuracy in primary role identification.

Next section will provide further details on these activities and the results obtained.

2 Multimodal Fake News Detection

This study (also described in [5]) comprehensively evaluates the Themis architecture, developed as part of the contributions to WP1, in the context of multimodal fake news detection on two distinct public datasets: Fakeddit and ReCoVery. To enhance model performance, we systematically investigate various customizations of Themis, including the integration of Low-Rank Adaptation, diverse data augmentation techniques, and multiple configurations, employing the TinyLlama Large Language Model and CLIP ViT image encoders while tuning key parameters to optimize results.

2.1 Datasets

The two datasets used to evaluate the Themis architecture are Fakeddit and ReCOVery, both of which are multimodal (text and images) but have very different

Fakeddit This is a large-scale dataset designed specifically for detecting fake news across different data modalities. The data comes from the social network Reddit. It includes metadata such as post titles, image URLs, and annotations for classification. It provides labels for 2-, 3-, or 6-class analysis. The 6-class classification distinguishes between: True, Satire, Misleading Content, Impostor Content (bot-generated), Incorrect Connection (mismatched text and image), and Manipulated Content. It consists mainly of short titles and slang, with an average of only 9 tokens per instance. Although the original dataset exceeds one million posts, the researchers extracted a sample of 10,000 posts to test the model in data-scarce scenarios.

ReCOVery This dataset was created specifically to combat misinformation related to the COVID-19 pandemic. The data comes from a variety of sources, including newspaper articles (such as The New York Times) and Twitter posts (X). Each record is very rich and includes the URL, publisher, author, title, associated image, political orientation, and the full text of the article. The news items are classified binarily as reliable or unreliable. Unlike Fakeddit, it contains full articles with an average of 900 tokens per instance, offering much greater contextual depth. A total of 1,409 articles (excluding tweets) were used, revealing a significant imbalance between reliable (1,211) and unreliable (198) news items, which is why data augmentation techniques were applied.

The main difference lies in the content: while Fakeddit covers a wide range of topics (from basketball to politics) with very short texts, ReCOVery is thematically focused on COVID-19 and features long, structured texts, allowing the model to perform better thanks to the consistency of linguistic patterns.

2.2 Data augmentation

Data augmentation techniques, such as TSIT and MixGen, were incorporated into the Themis architecture to address data scarcity and improve model robustness, but their impact varies significantly depending on the dataset used.

2.2.1 TSIT (Text Synonyms and Image Transformations)

In the TSIT data augmentation strategy, the selection of synonyms using WordNet follows a structured, multi-step process designed to ensure that the original meaning of the text remains unchanged. The sentences of the original text are first broken down into individual words. Next, lemmatization is applied, a technique that reduces each word to its base form or root (the "lemma"), facilitating accurate retrieval from the lexical database. Not all words in the text are replaced. The system selects a random portion of the terms in the sentence to proceed with the replacement. For the selected words, synonyms are extracted directly from WordNet, a widely used English lexical database in Natural Language Processing (NLP), specifically utilizing the Python library NLTK. The main objective of this method is to create syntactic variants (i.e., different ways of writing the same thing) that preserve the core meaning of the information, ensuring that a "fake" news story remains as such even after being reworded.

As for the augmentation of images, they undergo a series of standard transformations designed to modify their visual characteristics in order to increase the diversity of the dataset without altering the meaning of the original content. The specific transformations applied include flipping (mirror reflection), rotation, adjustment of brightness and

contrast, application of Gaussian blur, and resizing. Even in this case the goal is to generate multiple versions of the images to mitigate model overfitting and increase data variability, especially in resource-constrained contexts such as the ReCOVery dataset. Each augmented image is finally associated with its corresponding text record to ensure that every entry in the final dataset is unique.

Results on ReCOVery : This was the most successful strategy, propelling the model to the top of the performance rankings with an accuracy of 0.975 and a Macro-F1 score of 0.948. This success is due to the severe data scarcity in this dataset, which makes generating new instances crucial to prevent overfitting.

Results on Fakeddit : It showed marginal improvements over the standard model (Acc: 0.780 vs 0.777), but was outperformed by configurations that used only LoRA and token merging.

2.2.2 MixGen

MixGen is a multimodal data augmentation technique designed to increase the number of training examples by creating new records from the combination of two existing records. Its operation is based on two distinct processes for images and text:

- Visual Component (Images): The new visual record is generated through linear interpolation between the pixels located in the same position in the two source images. In the cited study, a factor of λ=0.5 was used, meaning that the resulting image is an equal blend of the two originals.
- Text Component: For the text portion, the process is simpler and consists of concatenating the two original texts into a single block.

Results on ReCOVery It produced significant improvements, achieving an accuracy of 0.940 (with LoRA dropout 0.4), but still fell short of the results obtained with TSIT.

Results on Fakeddit In this case, MixGen performed worse than the standard model, dropping to an accuracy of 0.769.

2.3 Discussion

Based on the experimental results, TSIT consistently outperformed both the baseline (standard) model and the MixGen technique on both tested datasets. While MixGen showed significant fluctuations in results, sometimes even performing worse than the model without data augmentation, TSIT proved to be a more reliable strategy, where the generation of synonyms and image transformations helped prevent overfitting, expanding the dataset's variability more effectively than the record merging performed by MixGen. Furthermore, TSIT operates by creating syntactic variations of the text and visual modifications that preserve the semantic core of the original instance. In contrast, MixGen creates new records by linearly combining two different images and concatenating their texts, a process that can introduce greater ambiguity or noise into the data.

The influence of Reddit slang (characteristic of the Fakeddit dataset) on the learning of the Themis model is primarily linked to the brevity of the text and the lack of contextual depth compared to more structured datasets. The Reddit posts in the Fakeddit dataset consist mainly of short, slang-filled sentences averaging just 9 tokens per instance, whereas the ReCOVery dataset contains full articles averaging 900 tokens. This huge difference means the model has much less textual material to work with. Furthermore, Fakeddit covers an extremely broad and diverse range of topics (such as Donald Trump, the International Space Station, or Christianity), making it difficult to identify recurring linguistic patterns. In contrast, ReCOVery is thematically focused on COVID-19, providing the model with clearer and more consistent patterns to learn. In the study, only the post titles were used for the Fakeddit dataset, omitting user comments and other metadata that could have provided additional context for interpreting slang. The combination of extremely short texts and colloquial language limits the Large Language Model's (TinyLlama) ability to establish strong semantic relationships, leading to lower performance (Macro-F1 of 0.80) compared to that achieved with ReCOVery articles (Macro-F1 of 0.948)

3 Comparison of Multimodal Models for Manipulation and Fake News Detection

With the increasing spread of manipulated content and fake news across online platforms, understanding how artificial intelligence can automatically detect such information has become a crucial research challenge. Social networks like Reddit, X, and various news-sharing platforms represent major sources of information, but also fertile ground for misinformation and persuasive manipulation. A study was conducted to explore and compare the capabilities of two state-of-the-art multimodal architectures, Themis and HAMMER, in detecting deceptive or manipulative content that combines both textual and visual elements.

HAMMER [7] adopts a complex hierarchical design. It employs separate text and image encoders whose outputs are aligned through a contrastive-aware module that brings semantically related representations closer. A hierarchical cross-modal transformer then performs multi-level attention to reason over fine-grained relationships between visual and textual cues, enabling both classification and manipulation localization.

To enhance cross-modal consistency, both architectures were extended with ALBEF (Align Before Fuse) [3], a pre-trained vision-language model that improves image-text alignment through contrastive and matching objectives using large-scale datasets. The integration of ALBEF into Themis and HAMMER was designed to evaluate whether pre-aligned multimodal features could improve the models' robustness and accuracy on downstream manipulation detection tasks.

The experimental evaluation was conducted on three multimodal datasets:

- **DGM^4** [7]. This dataset provides a balanced distribution between real and manipulated examples, making it particularly suitable for evaluating binary multimodal classifiers such as Themis and HAMMER.

- **Fakeddit** [6]. The 2-way version of the dataset was employed, using only binary labels to align with the classification objectives of both models. Fakeddit's diversity

of content and topics makes it a challenging benchmark for testing generalization and robustness.

- **Multi Fake Detective (MFD)** [4]. This more recent dataset aggregates multimodal fake news from multiple online domains, including news outlets, social media, and fact-checking archives. It contains a large collection of memes and posts, each consisting of an image, a textual component, and metadata about the source. Labels indicate whether the content is **Manipulated** or **Authentic**, with some instances involving partial manipulations where only one modality is deceptive. MFD is particularly useful for testing cross-domain generalization due to its heterogeneous nature and inclusion of both linguistic and visual nuances of manipulation.

The datasets were used to assess both in-distribution performance and cross-dataset generalization, simulating realistic misinformation scenarios. Each model was trained on progressively larger subsets (5K to 30K plus the full training set) to evaluate scalability and sample efficiency. Performance was measured using standard quantitative metrics such as Accuracy, AUC, and Equal Error Rate (EER), along with computational indicators such as Training Time, Inference Time, and Inference Time per Instance. The results revealed complementary strengths: HAMMER achieved strong performance due to its deep reasoning layers, while Themis demonstrated greater efficiency and faster convergence. The integration of ALBEF led to consistent improvements in both architectures, particularly in terms of AUC and generalization. The best results for each dataset are the following:

DGM⁴						
Model	**Accuracy**	**AUC**	**EER**	**Training Time (min)**	**Inference Time (min)**	**Inference Per Instance (ms)**
Themis	0.734 ± 0.015	0.840 ± 0.023	0.248 ± 0.011	**7200.03**	42.55	72.32
HAMMER	**0.856**	**0.925**	**0.148**	-	29.42	34.77

Fakeddit						
Model	**Accuracy**	**AUC**	**EER**	**Training Time (min)**	**Inference Time (min)**	**Inference Per Instance (ms)**
Themis	$\mathbf{0.838 \pm 0.005}$	$\mathbf{0.913 \pm 0.010}$	0.167 ± 0.009	**424.23**	8.58	64.30
HAMMER	0.834 ± 0.012	0.910 ± 0.013	$\mathbf{0.165 \pm 0.012}$	2499.27	**4.97**	**37.27**

MFD						
Model	**Accuracy**	**AUC**	**EER**	**Training Time (min)**	**Inference Time (min)**	**Inference Per Instance (ms)**
Themis	$\mathbf{0.675 \pm 0.039}$	$\mathbf{0.655 \pm 0.014}$	$\mathbf{0.384 \pm 0.008}$	**16.67**	**0.07**	**46.67**
HAMMER	0.670 ± 0.032	0.580 ± 0.017	0.430 ± 0.032	91.17	0.35	233.33

Overall, this comparative analysis provides a clearer understanding of how model complexity, multimodal alignment, and pre-training strategies influence the performance of manipulation and fake news detection systems.

4 Data-Centric Propaganda Analysis

Given the rapid proliferation of persuasive and manipulative narratives in online media, there is a growing need to understand how artificial intelligence models can effectively identify propaganda across different modalities. Despite significant progress, research in this area remains fragmented, with diverse approaches, datasets, and evaluation standards. For this reason, a survey was conducted to provide a unified overview of the existing methodologies, highlighting their strengths, limitations, and open challenges in

textual and multimodal propaganda detection. It presents a comprehensive and systematic examination of current methodologies for propaganda detection across textual and multimodal data. It organizes existing research into three main subtasks, each addressing a different level of granularity in identifying propagandistic content:

- **Span identification:** the detection and labeling of specific text fragments (tokens or phrases) that convey propagandistic techniques.
- **Binary classification:** the prediction of whether an entire text unit (sentence, paragraph, or document) contains propaganda or not.
- **Multi-label classification:** the assignment of one or more propaganda techniques to a given text from a predefined set of labels.

For each subtask, the study reviews datasets, models, and experimental strategies, highlighting both the technical progress achieved and the persistent limitations in the field.

The analysis reveals a clear dominance of transformer-based architectures, particularly fine-tuned models such as BERT, RoBERTa, and XLM-RoBERTa, which consistently outperform large language models such as GPT-4 in zero- and few-shot settings. At the same time, it shows that data imbalance, domain shift, and the lack of multilingual and multimodal resources remain major challenges. Span identification emerges as the most complex task, often suffering from scarce annotated data and fine-grained label distributions. Binary classification tasks tend to be more stable but struggle with generalization across topics and languages. Multi-label classification, while more realistic, exposes the limits of current architectures in modeling label correlations and subtle rhetorical cues.

From a critical standpoint, the survey emphasizes the fragmentation of current research — inconsistent taxonomies, heterogeneous annotation schemes, and variable evaluation setups hinder comparability and reproducibility. Moreover, despite the rapid evolution of deep learning approaches, interpretability and explainability remain underexplored, though recent works integrating rationale generation represent a promising trend.

The work [8] conducted a focused analysis of the main datasets for span-level propaganda detection, identifying ten corpora released between 2019 and 2025 and selected based on relevance, public availability, and richness of annotation schemes. These include well-established benchmarks such as PTC and SemEval-2020 Task 11, built on English news articles with fine-grained annotations of propaganda techniques and character-level offsets; the multilingual extension introduced by SemEval-2023 Task 3, covering six languages and broadening thematic scope; Arabic-focused resources such as WANLP, ArAIEval, and ArMPro, adopting annotation schemes with 21–23 techniques across news articles and tweets; and more context-specific datasets such as ZenPropaganda (Russian COVID-19-related media), BanMANI (Bangla manipulated posts with binary and span annotations), the English–Roman Urdu code-switched corpus, and ManiTweet (tweet–article pairs annotated for manipulative spans). The comparative analysis examined dataset size, label distribution, domain (news vs. social media), linguistic coverage, and label granularity, revealing substantial heterogeneity in annotation taxonomies, technique definitions, and structural design, as well as a predominance of the news domain over social media contexts. A pervasive issue across corpora is class imbalance, with frequently occurring techniques such as `Loaded Language` and `Name Calling` heavily overrepresented compared to rarer strategies, leading to learning instability and evaluation challenges. The review of studies employing these benchmarks further shows that, despite the adoption

of transformer-based architectures and Large Language Models, performance remains sensitive to imbalance, annotation noise, cross-lingual transfer, and cross-domain generalization, particularly in fine-grained and multi-label settings. Overall, the progress in propaganda detection does not rely solely on architectural innovation, but requires a data-centric perspective emphasizing clearer and more consistent guidelines, rigorous annotation validation, improved class balancing, and continuous dataset maintenance, in order to ensure robustness, comparability, and alignment with the evolving nature of online propaganda strategies.

5 Sub-Narrative Classification

The paper [1] presents the system developed by the team *iLostTheCode* for Subtask 2 of SemEval-2025 Task 10, which focuses on the multilevel classification of narratives in multilingual news articles. The task consists of assigning to each article one or more sub-narratives drawn from a two-level hierarchical taxonomy, including general narratives and their corresponding specific sub-categories. The problem is formulated as both multi-label and multi-class and involves five languages, namely Bulgarian, English, Portuguese, Hindi, and Russian, across articles mainly related to two broad topics: the Russia–Ukraine war and climate change. The official evaluation primarily relies on the Samples F1 metric, which requires the correct identification of both the narrative and its associated sub-narrative. The proposed approach follows a bottom-up strategy and combines contextual embeddings generated by multiple pre-trained BERT-based models with a simple neural network for final classification. During preprocessing, each article is split into sentences, which are treated as independent units in order to comply with Transformer token limits; all sentences belonging to the same article share the document's global labels. Languages other than English are automatically translated into English prior to processing. For each sentence, the CLS token embedding is extracted from several models that were previously fine-tuned on different tasks but not retrained on the competition dataset, including models for emotion classification, natural language inference, and named entity recognition. Configurations both with and without a contextual window including the two preceding sentences were experimented with. The embeddings produced by the different models are concatenated into a single vector that serves as input to a neural network composed of a ReLU layer, a dropout layer, and a final classification layer with either sigmoid or softmax activation, depending on the language and experimental configuration. Training is carried out in two phases: an initial phase using the Adam optimizer for rapid convergence, followed by a second phase using stochastic gradient descent with momentum to improve generalization and reduce overfitting, with early stopping based on validation loss. Sentence-level probabilities are aggregated at the article level through summation and normalization, while final class selection is performed via thresholding optimized on the development set, a stage that the authors identify as particularly critical for overall performance. The dataset includes approximately 400–450 annotated articles per language, with a smaller number for Russian, and exhibits marked class imbalance, with a high prevalence of the "Other" category. Official results show varying performance across languages, with the best outcome achieved in Russian and the lowest in Hindi, while in the remaining languages the system generally ranks above the average of participants. Subsequent post-task experiments indicate that simple adjustments of threshold values can yield measurable improvements without retraining the model. Error analysis through confusion matrices reveals a tendency to produce false positives in more frequent classes

and difficulty in recognizing rare classes, confirming the impact of dataset imbalance. Possible future improvements, including the adoption of separate classifiers for different thematic domains, multilingual integration without prior translation, and the exploration of top-down hierarchical strategies, while noting that performance at the general narrative level is not particularly critical compared to that at the sub-narrative level.

6 Entity Framing and Role Labeling

In this work [2] presents a system developed for SemEval-2025 Task 10, Subtask 1, focused on multilingual Entity Framing, namely the automatic assignment of main roles (protagonist, antagonist, innocent) and fine-grained sub-roles to entities mentioned in news articles primarily dealing with the Russia–Ukraine war and climate change, across five languages (English, Bulgarian, Hindi, Portuguese, and Russian). The task is formulated as a span-based multi-class and multi-label classification problem, evaluated mainly through the Exact Match Ratio (EMR), a particularly strict metric requiring the exact correspondence of all assigned labels. The proposed approach relies on parameter-efficient fine-tuning of Meta-Llama-3-8B-Instruct using QLoRA, which combines 4-bit quantization and Low-Rank Adaptation to reduce computational costs while maintaining competitive performance. The system pipeline consists of three main stages: data pre-processing, prompt design, and fine-tuning. During pre-processing, articles are segmented into sentences, a local context window around each entity is extracted, non-English texts are automatically translated into English, and the training data is augmented to mitigate class imbalance by generating paraphrased examples with an LLM and filtering them to preserve entity consistency and semantic coherence. Three prompting strategies were explored: S1, a single prompt containing all associated sub-roles; S2, separate prompts for each fine-grained role; S3, a mixed approach. Results on the development set show that S2 provides the best overall balance, significantly improving EMR and micro-F1 in several languages, while S1 struggles with simultaneous multi-label prediction and S3 does not consistently outperform S2. On the official test set, the system achieves an average main-role accuracy of approximately 0.84 and an average EMR of 0.41 for fine-grained roles, with performance varying across languages. The error analysis conducted on the Hindi development subset highlights persistent difficulties with rare sub-roles, e.g. Guardian, Traitor, Rebel, Spy and frequent confusions between semantically related categories such as Exploited and Virtuous, sometimes affecting main-role classification as well; additional challenges arise from passive constructions, ironic expressions, and limited contextual information. Overall, while the model demonstrates solid capability in identifying main roles, it encounters greater difficulty in achieving exact multi-label sub-role matching, suggesting that future improvements may involve refined prompting strategies and alternative fine-tuning configurations to enhance fine-grained role discrimination in multilingual settings.

References

Publications Acknowledging the DEMON Project

[1] Lorenzo Vittorio Concas, Manuela Sanguinetti, and Maurizio Atzori. “iLostTheCode at SemEval-2025 Task 10: Bottom-up Multilevel Classification of Narrative Tax-

onomies". In: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Ed. by Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 607–616. ISBN: 979-8-89176-273-2. URL: https://aclanthology.org/2025.semeval-1.85/.

[2] Matteo Fenu, Manuela Sanguinetti, and Maurizio Atzori. "DEMON at SemEval-2025 Task 10: Fine-tuning LLaMA-3 for Multilingual Entity Framing". In: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Ed. by Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1456–1464. ISBN: 979-8-89176-273-2. URL: https://aclanthology.org/2025.semeval-1.192/.

[5] Davide Antonio Mura, Marco Usai, Andrea Loddo, Manuela Sanguinetti, Luca Zedda, Cecilia Di Ruberto, and Maurizio Atzori. "Is it fake or not? A comprehensive approach for multimodal fake news detection". In: *Online Soc. Networks Media* 47 (2025), p. 100314. DOI: 10.1016/J.OSNEM.2025.100314. URL: https://doi.org/10.1016/j.osnem.2025.100314.

[8] Marco Usai, Davide Antonio Mura, Andrea Loddo, Manuela Sanguinetti, Luca Zedda, Cecilia Di Ruberto, and Maurizio Atzori. "Exploring the Dataset Landscape for Automated Propaganda Detection: A Data-Centric Insight". In: *Proceedings of the 4th Italian Conference on Big Data and Data Science (ITADATA 2025), Turin, Italy, September 9-11, 2025*. Vol. 4152. CEUR Workshop Proceedings. CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4152/short71.pdf.

Other References

[3] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. "Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation". In: *arXiv preprint arXiv:2107.07651* (2021). DOI: 10.48550/arXiv.2107.07651. URL: https://arxiv.org/abs/2107.07651.

[4] Multi-Fake Detective Project Team. *Multi-Fake Detective Dataset*. https://sites.google.com/unipi.it/multi-fake-detective/data. Accessed: 2025-07-18. 2024.

[6] Kai Nakamura, Sharon Levy, and William Yang Wang. "Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection". In: *arXiv preprint arXiv:1911.03854* (2019).

[7] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. "Detecting and Grounding Multi-Modal Media Manipulation". In: *arXiv preprint arXiv:2304.02556* (Apr. 2023). DOI: 10.48550/ARXIV.2304.02556. URL: https://arxiv.org/abs/2304.02556.