SoMeLVLM: A Large Vision Language Model for Social Media Processing

Anonymous ACL submission

Abstract

The growth of social media, characterized by its multimodal nature, has led to the emergence of diverse phenomena and challenges, which calls for an effective approach to uniformly solve automated tasks. The powerful Large Vision Language Models make it possible to handle a variety of tasks simultaneously, but even with carefully designed prompting methods, the general domain models often fall short in aligning with the unique speaking style and context of social media tasks. In this paper, we introduce a Large Vision Language Model for Social Media Processing (SoMeLVLM), which is a cognitive framework equipped with five key capabilities including *knowledge & comprehension*, *application*, *analysis*, *evaluation*, and *creation*. SoMeLVLM is designed to understand and generate realistic social media behavior. We have developed a 654k multimodal social media instruction-tuning dataset to support our cognitive framework and fine-tune our model. Our experiments demonstrate that SoMeLVLM achieves state-of-the-art performance in multiple social media tasks. Further analysis shows its significant advantages over baselines in terms of cognitive abilities.

Figure 1: An illustration showing that general domain large language models encounter troubles in (a) social multimedia understanding, (b) informal language understanding, and (c) complex cognitive demands in social media tasks.

1 Introduction

Online social media platforms have been generating an abundance of textual and visual content, offering insights into how individuals communicate, interact, and express themselves. With the advent of communication technology, social media is receiving growing attention as more and more users are active in communities of various topics and interests, which is becoming an important research object as well as a valuable data resource for Computational Social Science (CSS) research (Lazer et al., 2020). Consequently, automated tasks like sentiment analysis (Saravia et al., 2018) and misinformation detection (Gabriel et al., 2022) have emerged to help researchers understand social media users and optimize online communities.

Recently, Large Language Models (LLMs) and Large Vision Language Models (LVLM) (OpenAI, 2023; Zhang et al., 2023; Touvron et al., 2023b; Chiang et al., 2023; Lyu et al., 2023) have demonstrated their immense capabilities and have offered an effective way to handle automated tasks through prompt engineering. However, research has shown that these generic large models even with extensive prompting practices and evaluations cannot completely replace the traditional research pipeline for CSS, particularly in social media studies (Ziems et al., 2023). As illustrated in Figure 1, we discover three major challenges faced by general domain

models in addressing the nuances of social media:

Limitations in social multimedia understanding. General domain LLMs or LVLMs tend to focus more on text over other modalities, which is **not** consistent with real-world user habits on social media (Liu et al., 2023; Li et al., 2023b; Dai et al., 2023; Zhu et al., 2023). Social media tasks often require fine-grained recognition ability to combine captions and images from a single post and synthesize the user's intention. Genereal domain large models may not possess this level of nuanced multimodal understanding, as shown in Figure 1 (a).

Challenges in informal language understanding. There is a huge gap between the informal speaking style prevalent on social media and the formal language used in other contexts. As a result, general domain LLMs and LVLMs fall short in recognizing sentiment, humor, figurative language, and other related concepts when the sentences are expressed casually. The example shown in Figure 1 (b) demonstrates that the model cannot recognize the wordplay "banded" in the user's post.

Complex cognitive demands in social media tasks. Social media tasks often involve multiple objectives to address high-level social demands that require a combination of complex cognitive abilities and information-processing levels. For instance, the detoxifying task illustrated in Figure 1 (c), involves both hate speech detection and content rewriting. However, the models without these abilities struggle to comprehensively address these aspects, resulting in less than satisfactory outputs.

Therefore, to overcome these limitations of the simple prompting strategies and shed light on the investigation of *"how LLMs produce new CSS paradigms built on the multipurpose capabilities of LLMs over the long term"* (Ziems et al., 2023), we propose **SoMeLVLM**, a large vision language model tailored for social media processing via extensive and comprehensive supervised fine-tuning. In particular, we establish a solid theoretical foundation. We categorize the tasks concerning social media systematically and build a cognitive pyramid based on Bloom's Taxonomy (Bloom and Krathwohl, 1956), including cognitive levels of *Knowledge & Comprehension*, *Application*, *Analysis*, *Evaluation*, and *Creation*. These cognitive abilities are derived from different types of users on social media and represent different levels of demands for information processing.

To infuse our model with cognitive abilities, we have curated a large-scale multimodal dataset comprising a total of 654k instances of plain-textual and multimodal data. We then formulate these data into instruction data formats by designing multiple instructional prompts for each task-related subset, covering 12 tasks in total including *emotion*, *humor*, *figurative language*, *hate speech & toxicity*, *ideology & stance*, *misinformation*, *trustworthiness & social bias*, *social factors*, *detoxifying content*, *depolarizing language invert opinion*, and *reverse ideology*. Both classification and generative tasks are included in our dataset.

We apply instruction tuning to our model in two steps. The base language model is tuned initially using textual instruction data, and then a connection module between the vision encoder and the base language model is tuned using multimodal data for advanced cognitive abilities.

We have conducted both in-domain and out-of-distribution tests on our model and evaluated the performance at both task and cognitive ability levels. The results show that our model effectively overcomes these limitations and achieves state-of-the-art performance in various social media tasks.

To summarize, the main contributions of our paper are as follows:

- We propose a large vision language model specifically tailored for social media contexts, capable of delivering high-quality text classification and interpretation under zero-shot conditions, fundamentally simplifying the research workflow in computational social science and improving overall reliability.
- We construct a comprehensive social media framework by combining cognitive abilities with traditional social media tasks to support different levels of demands in information processing.
- We contribute to a large-scale, high-quality multimodal social media dataset, encompassing both pure text and multimodal formats, with data from both open-source and self-collected sources, formatted into diverse instruction-tuning formats.

2 Related Works

2.1 Computational Social Science

As an interdisciplinary field, Computational Social Science (Lazer et al., 2020; Edelmann et al., 2020) leverages computational methods to analyze vast datasets, encompassing data from everyday conversations, documents, and books, as well as **social**

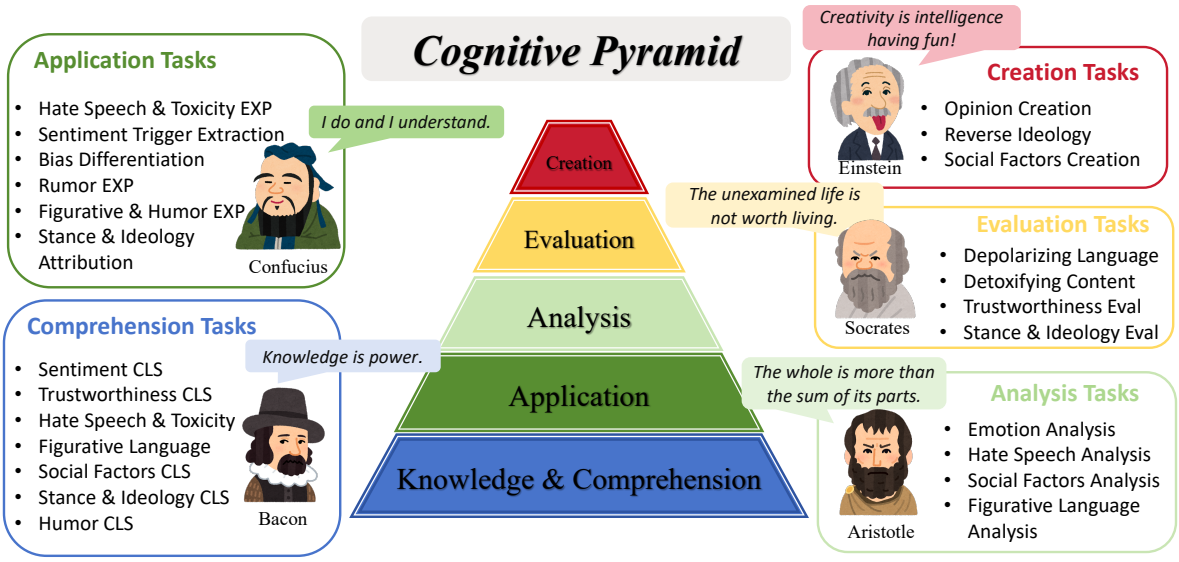

Figure 2: An illustration of the Social Media Cognitive Framework.

media content, to scientifically study linguistic behaviors and social phenomena (Lazer et al., 2009; Keuschnigg et al., 2018).

The rise of the Internet has made online interactions a fundamental part of daily life (Golder and Macy, 2014), providing invaluable resources for Computational Social Science (Shah et al., 2015), and paving the way for advancements in social linguistic analysis, such as humor detection (Holton and Lewis, 2011), stance detection (ALDayel and Magdy, 2021), detection of figurative language (Reyes et al., 2012), and sentiment analysis (Neri et al., 2012). Furthermore, it provides guidance for predicting social phenomena, such as fake news detection (Shu et al., 2017), the recognition of hate speech (Mondal et al., 2017) and the prediction of ideologies (Mou et al., 2023), contributing to a deeper understanding of online and offline social dynamics.

2.2 Large Vision Language Model

The exceptional text understanding and generation capabilities demonstrated by large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023a; Zhang et al., 2023; Chiang et al., 2023; Lyu et al., 2023) have garnered attention across various fields. To further enhance the capability of instruction understanding and generalization ability on unseen datasets, researchers have employed instruction tuning (Wei et al., 2022; Chung et al., 2022) on LLMs. This approach is capable of augmenting LLMs' comprehension of language within specific domains (Bao et al., 2023; Yue et al., 2023; Chen et al., 2023), such as medicine, law, and finance, thereby enhancing performance on related tasks.

By integrating the visual encoders (Radford et al., 2021; Fang et al., 2023) and large language models through linear projection (Tsimpoukelli et al., 2021), Q-former (Li et al., 2023b) or cross-attention layers (Alayrac et al., 2022), LVLMs is capable of addressing a wide range of multimodal tasks. Researchers have also employed instruction tuning on LVLMs, including multitask learning (Cho et al., 2021), additional visual components (Li et al., 2023b; Alayrac et al., 2022), and instruction-aware components (Dai et al., 2023). By adopting such an approach, there has indeed been an enhancement in the models' zero-shot generalization capabilities.

3 Social Media Cognitive Framework

In this section, we will present the design of the cognitive pyramid for SoMeLVLM.

3.1 Framework Design

To construct a large vision language model capable of understanding and creating multimodal content on social media, we consider concepts from cognitive teaching methods and build a comprehensive multimodal social media cognitive framework, as depicted in Figure 2. We begin by designing a cognitive pyramid according to Bloom's Taxonomy (Bloom and Krathwohl, 1956), which is a classic teaching theory proposed by Benjamin Bloom in 1956. The pyramid contains five cognitive lev-

Level	Category	SFT DataSize	Eval Datasize	Total
Knowledge & Comprehension	Emotion	63.8k	6.5k	70.3k
	Humor	18.0k	8.3k	26.3k
	Figurative Language	12.5k	4.6k	17.1k
	Misinformation	30.4k	2.5k	32.9k
	Hate Speech & Toxicity	56.5k	7.7k	64.2k
	Ideology & Stance	25.3k	3.8k	29.1k
	Trustworthiness & Social Bias	11.0k	3.2k	14.2k
	Social Factors	55.2k	3.5k	58.7k
Application	Emotion	20.0k	5.0k	25.0k
	Humor	15.0k	6.1k	21.1k
	Hate Speech & Toxicity	29.6k	16.2k	45.8k
	Ideology & Stance	4.3k	1.0k	5.3k
	Trustworthiness & Social Bias	30.0k	-	30.0k
	Social Factors	49.0k	1.0k	50.0k
Analysis	Figurative Language	30.0k	2.2k	32.2k
	Emotion	18.8k	1.5k	20.3k
	Hate Speech & Toxicity	12.3k	1.5k	13.8k
	Social Factors	14.5k	0.5k	15.0k
Evaluation	Ideology & Stance	1.3k	0.3k	1.6k
	Misinformation	8.0k	0.5k	8.5k
	Trustworthiness & Social Bias	-	0.9k	0.9k
	Detoxifying Content	25.0k	9.9k	34.9k
	Depolarizing Language	4.3k	1.0k	5.3k
Creation	Invert Opinion	1.0k	-	1.0k
	Reverse Ideology	4.3k	1.0k	5.3k
	Social Factors	24.5k	0.5k	25.0k
Total		564.6k	89.2k	**653.8k**

Table 1: Composition of data for different cognitive levels

els: *Knowledge & Comprehension*, *Application*, *Analysis*, *Evaluation*, and *Creation*.

We then construct the instruction-tuning data for these five cognitive levels, which is a combination of existing datasets and data collected from social media, resulting in a total of **654k** instruction pairs. The relation between cognitive levels and different tasks and data statistics are presented in Table 1. Each data instance is structured into `text_input`, `text_output`, and `image` if it is multimodal, aligning with the format used in Blip2 (Li et al., 2023b). To ensure the quality of the instruction pairs, we manually design five prompts for *each* dataset. Detailed examples of both plain text and multimodal types are provided in Appendix A.2.

3.2 Knowledge & Comprehension Level

The Knowledge & Comprehension level means to recall and understand basic facts. It represents a basic cognitive ability in our framework, which is also the foundation of other higher-level cognitive abilities. Tremendous amounts of concepts are learned via real-world social media data at this level to help the model recognize the content on social media.

Specifically, the instruction construction of this level consists of various classification tasks within the context of social media, featuring a basic understanding without deeper analysis. We have collected a comprehensive collection of open-source datasets annotated by experts in areas such as *Emotion*, *Humor*, *Figurative Language*, *Misinformation*, *Hate speech & Toxicity*, *Ideology Stance*, *Trustworthiness & Social Bias*, and *Social Factors*. These datasets are structured into question-answering formats, prompting the language model to recognize and categorize these concepts from samples in both textual and multimodal datasets. For binary classification or pairwise choices, a true-or-false question format is applied. For multi-classification, the choices include the entire label space containing up to six candidate answers.

3.3 Application Level

The Application level means to use the information in new situations, which is related to active involvement in social media. Concepts learned at the former level are used at the application level to explain the phenomena on social media. Consequently, the instruction construction is to make accurate interpretations based on the given ground truth over various social media domains, implying an understanding of the reasons behind the labels.

Given the original ground truth within the datasets annotated by experts, the `text_output` of the instruction pair is formulated by appending a concise explanation after the ground truth. Data following the above steps are formulated into tasks including *Emotion Trigger Extraction*, and Interpretation of *Humor*, *Hate Speech*, *Ideology & Stance*, *Trustworthiness*, and *Social Factors*. For unlabeled data we collect from social media, the ground truth labels are designed as hashtags, personalities, and fields that are closely related to social media. The generated labels along with the explanation are generated by the powerful language model like GPT-4 in advance. To put it briefly, the primary characteristic of the application level is: **given existing labels**, it enables the model to generate corresponding explanations.

3.4 Analysis Level

The Analysis level means to draw connections among ideas, which is similar to the application level in that it is a second process based on the concepts learned at the Knowledge & Comprehension level. The analysis level requires the model to analyze the label and furnish the corresponding interpretations independently. This implies a higher order of capability, enabling it to navigate the rapidly evolving social media landscape.

We aim for the model to offer explanations **in the absence of ground truth labels** at this level. Given the original text or text-image pairs, we provide only the broad context necessary for the analysis of the model such as *Figurative Language Analysis*, *Emotion Analysis* and *Hate Speech Analysis*, and then let the model autonomously generate labels and corresponding explanations. For instance, we instruct the model to analyze the emotional connotation conveyed by the text (or image-text-pair) and elucidate the reasons thereof, while at the application level, we directly present the ground truth emotion and direct the model to analyze the causative factors inducing the said emotion. Therefore, to construct the instruction pairs, the datasets are formulated into a question-answer format, where the question is reformed into a more complex instruction while the answer is generated by GPT-4.

3.5 Evaluation Level

The Evaluation level represents the risk forecasting ability, which stands for assessing the probability or likelihood of potential social events and predicting collective trends. At the evaluation level, we pay special attention to the existing prejudices within the data and the abnormal behavior on social media and prompt the model to rewrite original texts or apply knowledge from other domains.

The construction of the data is divided into two aspects. Firstly, for texts that are labeled as containing Hate Speech, we undertake detoxification, and for texts labeled as Liberal or Conservative, we engage in depolarization. Secondly, for texts or text-image pairs labeled as Misinformation, we instruct the model to explain the underlying reasons. Ultimately, the composition of the data is presented in a question-answer format, where the question corresponds to the specific instruction, and the answer is generated by GPT-4.

3.6 Creation Level

The Creation level means to create reliable content related to social media, which is essential during the interaction with the content on social media. This level is considered to be the most complex level. We tackle this demand by setting *reverse* and *creation* tasks, respectively. In the *reverse* task, we require the model to generate opposing viewpoints based on a specified topic and text. In the *create* task, the task is formulated as the generation of new hashtags on social media.

In terms of instruction construction, regarding the *reverse* task, we formulate the question to prompt the model to generate opposing views on a specific topic, while selecting real statements that hold contrary opinions as the answer. As for the *create* task, we prompt GPT-4 to generate new hashtags related to specific texts, thereby producing question-answer pairs.

4 Experimental Setup

4.1 Data Split

After the data construction following the design in §3, we fine-tune our model using around 564k

Models	Hate Speech		Misinformation		Social Factors		Emotion		Ideology		Social Factors OOD	
	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc
InstructBlip$_V$	41.62	33.43	47.55	13.60	80.02	40.93	54.53	48.90	54.15	42.41	87.30	22.59
InstructBlip$_F$	50.40	48.43	80.78	79.00	81.33	73.57	58.90	57.80	53.69	45.57	98.31	83.95
Blip2	52.14	52.14	80.60	80.60	81.83	**80.89**	57.73	57.73	53.48	53.48	99.15	**95.69**
Llava	53.35	9.79	**84.67**	25.40	72.49	6.69	53.39	10.10	49.79	1.58	93.75	3.08
MiniGPT4	45.12	23.00	65.30	54.20	64.08	36.18	53.13	29.48	42.13	8.86	69.58	34.29
SoMeLVLM	**72.57**	**72.57**	82.60	**82.60**	**84.07**	67.33	**63.50**	**63.47**	**73.24**	**55.06**	**100.00**	61.11

Table 2: Main results of multimodal classification tasks. We report Acc (overall accuracy) and Acc* (accuracy in instruction-following outputs). The **bold** number represents the best results, and the underlined number represents the second-best results.

training data, which is labeled as *SFT* in Table 6. We then evaluate our SoMeLVLM across various aspects of social media, marked as *Eval*, including 14 multimodal datasets and 12 held-out plain text datasets, totaling around 89k data. The specific datasets corresponding to each task and the provided instructions are detailed in the Appendix A.1.

4.2 Baseline Models

For tasks involving plain text, we select Llama-2-7b-chat-hf(Touvron et al., 2023b), Vicuna-7b-v1.1 (Chiang et al., 2023), and ChatGLM2-6b (Zeng et al., 2022) as our baseline models.

For tasks containing images, we choose Blip2 (Li et al., 2023b), InstructBlip (both Vicuna-based and FlanT5xl-based) (Dai et al., 2023), Llava (Liu et al., 2023), and Minigpt4 (Zhu et al., 2023) as our baseline models.

4.3 Evaluation Metrics

For classification (CLS) tasks, we report the accuracy (Acc) of test results, which involves string matching after proper processing. Specifically, considering the zero-shot setting and the overall instruction-following ability of LVLMs, we report both the accuracy over the whole test set and the accuracy when only valid answers are counted (**Acc***). For generative (GEN) tasks, we report on automatic metrics such as **BLEU** and **ROUGE**. In addition, we employ GPT-4 as a grading assistant through specific prompts to evaluate the test outcomes (**GPT-Score**). In particular, we task GPT-4 with scoring the model's response on a scale from 0 to 5, where a higher score signifies greater consistency with the ground truth. These prompts can be found in Appendix A.2.

4.4 Implementation Details

For base language model tuning, we employ the QLoRA method (Dettmers et al., 2023) with FastChat (Zheng et al., 2023). To tune the connection module, we conduct our experiment following the method of LAVIS (Li et al., 2023a) and choose the connection module of `blip-vicuna-instruct` as the initial model. Accordingly, the base language model to be fine-tuned is assigned as Vicuna-7b-v1.1. The training and inference process is carried out on eight NVIDIA GeForce RTX3090 and eight RTX4090 GPUs. A mixed precision strategy is employed during the training stage due to the restriction of memory. The base language model is first trained for two epochs with plain text datasets, then the connection module is trained on multimodal datasets for three epochs. In the evaluation stage, we employ `gpt-4-preview-1106` to output the final score.

5 Results

5.1 In-Domain Evaluation

Given the limited availability of multimodal datasets for social media, we primarily carry out the evaluation of multimodal parts under an in-domain setting. We test our model on 11 datasets across five domains including hate speech, misinformation, social factors, emotion, and ideology. The overall results for classification tasks and generative tasks are shown in Table 2 and Table 3, respectively. SoMeLVLM has significantly surpassed the baseline LVLMs in all of the five domains in both classification and generative tasks, demonstrating its robust ability to handle a wide range of computational social science tasks.

Models	Metrics	Hate Speech	Misinformation	Social Factors	Emotion	Ideology	Social Factors OOD
InstructBlip$_V$	BLEU	0.65	1.09	6.21	0.85	0.60	1.14
	ROUGE	3.13	0.88	9.02	7.26	4.89	14.03
	GPT Score	1.83	2.84	1.46	1.96	1.61	2.07
InstructBlip$_F$	BLEU	0.24	0.05	1.16	0.28	0.78	1.51
	ROUGE	2.79	0.81	14.60	13.69	8.36	16.91
	GPT Score	2.11	2.85	2.12	3.02	1.62	2.16
Blip2	BLEU	0.62	0.02	0.76	0.16	0.25	0.65
	ROUGE	2.25	1.89	11.99	14.82	4.35	12.87
	GPT Score	1.86	2.72	1.89	3.08	2.34	1.61
Llava	BLEU	0.36	0.00	1.89	0.64	1.10	2.29
	ROUGE	4.52	0.01	12.80	5.74	8.73	20.10
	GPT Score	1.23	0.81	1.80	1.25	1.21	2.27
Minigpt4	BLEU	0.43	0.69	1.20	0.55	0.32	1.98
	ROUGE	8.84	12.15	17.20	10.81	12.68	20.73
	GPT Score	2.28	2.18	1.59	2.37	1.28	1.84
SoMeLVLM	BLEU	**31.04**	**24.06**	**14.49**	**37.65**	**24.08**	**10.18**
	ROUGE	**46.35**	**43.22**	**32.87**	**53.87**	**41.04**	**31.03**
	GPT Score	**3.21**	**2.94**	**2.86**	**3.53**	**3.39**	**3.45**

Table 3: Main results of multimodal generation tasks. We report BLEU-L, ROUGE-L, and GPT Score (0 to 5). The **bold** number represents the best results, and the underlined number represents the second-best results.

Models	Emotion	Humor	Figurative language	Misinfo	Hate Speech	Ideology	Trustworth	Social Factors
Vicuna	35.86	41.08	47.07	59.23	11.94	34.15	36.60	42.68
Llama2	40.54	**61.31**	53.77	41.11	12.84	37.77	59.21	31.61
ChatGLM2	41.20	36.94	52.05	47.21	14.67	30.07	**68.44**	48.23
SoMeLVLM	**80.66**	60.47	**61.70**	**70.38**	**22.20**	**45.23**	43.52	**55.39**

Table 4: Main result of plain text classification tasks under OOD settings; we report Accuracy for these tasks. The **bold** number represents the best results, and the underlined number represents the second-best results.

5.2 Out-of-Distribution Evaluation

For plain-text parts, we conduct Out-of-Distribution (OOD) evaluation in eleven distinct areas, encompassing emotion, humor, figurative language, hate speech, misinformation, ideology, trustworthiness, social factors, detoxifying content, depolarizing language, and reverse ideology. As shown in Table 4 and Table 5, SoMeLVLM achieves new zero-shot SOTA results on all aspects. The OOD evaluation of multimodal parts in the social factors domain involving three custom datasets is also reported as *Social Factor OOD* in Table 2 and Table 3, which is consistent with the results in the in-domain evaluation.

5.3 Results Analysis on Cognitive Abilities

We reform the above results according to the cognitive abilities mentioned in our framework. Specifically, we collect the in-domain performance of multimodal parts (using overall Acc performance) and the OOD performance of plain-text parts at the dataset level and categorize them into *Knowledge & Comprehension*, *Application*, *Analysis*, *Evaluation*, and *Creation*, five cognitive levels in total.

The reformed results are shown in Figure 3. Clearly, SoMeLVLM shows greater cognitive ability over baseline models in all of the cognitive levels. At the multimodal *Creation* level, all of the models perform poorly as they are required to generate three hashtags that best describe the post,

Models	Metrics	Emo	Humor	Figura	Hate	Ideol	Trust	Detoxify	Depolar	Rever
Vicuna	BLEU	7.97	10.49	8.03	7.01	9.36	9.70	10.43	22.31	33.40
	ROUGE	31.31	36.21	31.55	31.24	32.78	34.13	27.96	42.72	51.76
	GPT	3.23	3.24	2.57	3.63	3.41	3.13	2.50	3.26	2.98
Llama2	BLEU	4.25	6.36	10.39	1.79	4.75	4.73	1.31	8.40	20.54
	ROUGE	23.50	28.37	31.32	17.41	25.01	26.54	10.94	26.72	38.06
	GPT	2.99	2.48	2.73	1.94	2.78	2.82	1.14	2.21	2.04
ChatGLM2	BLEU	6.60	8.98	7.20	4.50	6.59	9.25	6.84	13.33	21.91
	ROUGE	29.47	34.49	29.07	28.05	29.94	34.35	23.92	35.66	42.27
	GPT	3.05	2.37	2.06	2.93	2.86	2.73	2.00	2.80	2.80
SoMeLVLM	BLEU	**26.96**	**13.81**	**23.77**	**17.24**	**14.60**	**12.37**	**27.13**	**23.54**	**44.09**
	ROUGE	**51.88**	**42.84**	**45.42**	**43.10**	**39.49**	**39.06**	**47.76**	**45.47**	**61.96**
	GPT	**3.63**	**3.38**	**3.02**	**3.64**	**3.43**	**3.59**	**2.89**	**3.28**	**3.41**

Table 5: Main result of plain text generative tasks under OOD settings; we report BLEU-L, ROUGE-L, and GPT Score (0 to 5) for these tasks (Hate, Ideol, Trust, Depolar, and Rever denote Hate Speech, Ideology & Stance, Trustworthiness, Depolarize Language, and Reverse Ideology, respectively.). The **bold** number represents the best results, and the underlined number represents the second-best results.

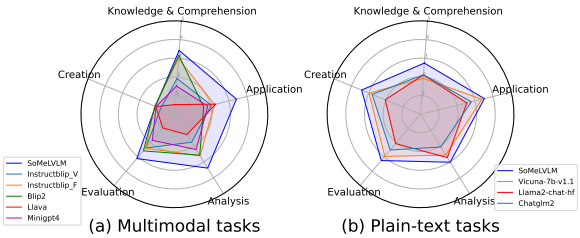

Figure 3: Cognitive abilities performances in (a) Multimodal tasks, and (b) Plain-text tasks.

which is not an easy task even for human beings.

5.4 Discussion on Instruction Following

We have noticed that the performance among LVLMs in Table 2 and Table 3 varies significantly, especially for Llava. The overall accuracy of Llava in the classification task is extremely poor, while the accuracy within the valid answer (namely, Acc*) looks good – even surpassing our model in the misinformation domain. This feeling of separation between Acc and Acc* results from the instruction-following ability of different base language models. When accompanied by the visual information provided by a visual encoder and connection module, base language models of LVLMs at 7b level show degeneration in following the output form according to the instructions. Specifically, in our baseline LVLMs, Llama-family (Vicuna-7b-v1.1 and Llama2) base models perform worse than the Flant5-family (Flant5xl) base model. Nevertheless, SoMeLVLM achieves overall the best performance even though we fine-tune it on Vicuna-7b-v1.1, which is the same as InstructBlip$_V$.

Research has found that the ability of instruction-following in LVLMs can be recovered under the few-shot settings (Li et al., 2023c). However in the CSS domain, especially in social media tasks, the zero-shot setting is more proper than a few-shot, as we hope to find a paradigm to handle these tasks automatedly. Besides, in this paper, we want to cultivate complicated cognitive abilities into our model instead of simply emphasizing instruction-following ability, which only belongs to the Knowledge & Comprehension level.

6 Conclusion

In our work, we introduce SoMeLVLM, a multimodal language model for social media processing, wherein we design five cognitive capabilities, each of which is mapped to various levels of social media tasks. Building on this, we collect related plain text and multimodal datasets and enhance the capabilities of vision-language models on relevant tasks through instruction tuning. Additionally, we construct an evaluation based on cognitive levels and test our model under zero-shot conditions, comparing it with other advanced LLMs and LVLMs. The experimental results thoroughly demonstrate the superiority of our model. Our work contributes to the computational social science field by providing methods for modeling and evaluating various tasks on social media and a large-scale, high-quality multimodal social media dataset.

Limitations

Our work currently focuses on English, and the performances shown in this paper may not be well reproduced in other languages. We are working on a multilingual dataset to improve the robustness under multilingual circumstances. On the other hand, these neologisms and phrases are often driven by specific cultures, communities, or events, and their meanings may vary across different groups. This suggests that our SoMeLVLM could exhibit interpretive biases towards these terms, especially in the absence of context.

Ethics Statement

The data used in this paper are from real users in diverse social media platforms, so the privacy problem is treated cautiously. The data from open-source datasets are safe as the sensitive information has already been masked. For the data we collect, we strictly follow the privacy policy of social media platforms and will carefully avoid personal information before we release our instruction dataset.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing Management*, 58(4):102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation.

Benjamin S. Bloom and David R. Krathwohl. 1956. *Taxonomy of educational objectives; the classification of educational goals by a committee of college and university examiners. Handbook I: Cognitive Domain*. Longmans, Green, New York, NY.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Conference on Empirical Methods in Natural Language Processing*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Julia Elliott Lucas Dixon Mark McDonald nithum Will Cukierski cjadams, Jeffrey Sorensen. 2017. Toxic comment classification challenge.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. Computational social science and sociology. *Annual Review of Sociology*, 46(1):61–81.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Scott A. Golder and Michael W. Macy. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40(1):129–152.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Felipe González-Pizarro and Savvas Zannettou. 2022. Understanding and detecting hateful content using contrastive learning.

Justin H Gross, Brice Acree, Yanchuan Sim, and Noah A Smith. 2013. Testing the etch-a-sketch hypothesis: a computational analysis of mitt romney's ideological makeover during the 2012 primary vs. general elections. In *APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting*.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avery Holton and Seth Lewis. 2011. Journalists, social media, and the use of humor on twitter. *Electronic Journal of Communication*, 21.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Marc Keuschnigg, Niclas Lovsjö, and Peter Hedström. 2018. Analytical sociology and computational social science. *Journal of Computational Science*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884.

David Lazer, Alex Pentland, L. Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, Jessica Fowler, and Myron Gutmann. 2009. Life in the network: The coming age of computational social science. 323.

David M. J. Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen

Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, Xuanjing Huang, and Zhongyu Wei. 2023c. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 85–94, New York, NY, USA. Association for Computing Machinery.

Xinyi Mou, Zhongyu Wei, Lei Chen, Shangyi Ning, Yancheng He, Changjian Jiang, and Xuanjing Huang. 2021. Align voting behavior with public statements for legislator representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1236–1246, Online. Association for Computational Linguistics.

Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuanjing Huang. 2023. UPPAM: A unified pre-training architecture for political actor modeling based on language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11996–12012, Toronto, Canada. Association for Computational Linguistics.

Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. 2012. Sentiment analysis on social media. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 919–926.

OpenAI. 2023. ChatGPT. `https://chat.openai.com/`. Accessed: 2024-02-03.

Francisco Manuel Rangel Pardo, Paolo Rosso, Manuel Montes y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *Conference and Labs of the Evaluation Forum*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowledge Engineering*, 74:1–12. Applications of Natural Language to Information Systems.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Dhavan V. Shah, Joseph N. Cappella, and W. Russell Neuman. 2015. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1):6–13.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Magdalena Wojcieszak, Andreu Casas, Xudong Yu, Jonathan Nagler, and Joshua A. Tucker. 2022. Most users do not follow political elites on twitter; those who do show overwhelming preferences for ideological congruity. *Science Advances*, 8(39):eabn9418.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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 open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Xiyuan Zhang, Xinyue Zhang, and Ying Yu. 2023. Chatglm-6b fine-tuning for cultural and creative products advertising words. pages 291–295.

Yunxiang Zhang and Xiaojun Wan. 2022. Mover: Mask, over-generate and rank for hyperbole generation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

A Supplementary on Data Collection and Processing

A.1 Datasets

Our datasets come from existing open-source datasets and the raw data we collect. Table 6 shows all datasets and their relations with cognitive modules and social media tasks. The categories of tasks has been expanded based on the foundation provided by SOCKET(Choi et al., 2023).

A.1.1 Existing Datasets

The following are open-source datasets categorized according to task:

Emotion Binary dataset for coarse-grained sentiment classification: Sentiment140 (Go et al., 2009); Multi-class dataset for fine-grained emotion classification: CARER (Saravia et al., 2018). MVSA_Single and MVSA_Multiple (Gomez et al., 2020), TumEmo (Yang et al., 2020).

Humor Binary datasets for humor classification: hahackathon (Meaney et al., 2021), reddit_jokes/puns/short_jokes (Weller and Seppi, 2019), humor-pairs (Hossain et al., 2020).

Figurative Language Binary datasets for coarse-grained figurative language classification: sar (Khodak et al., 2018); tweet_irony (Van Hee et al., 2018); a multi-class dataset for fine-grained figurative language classification: FLUTE (Chakrabarty et al., 2022).

Misinformation Binary datasets for misinformation classification: climate_change/cancer (Gabriel et al., 2022), FakeNewsNet (Shu et al., 2018).

Hate Speech & Toxicity Binary datasets for coarse-grained hate speech classification: implicit-hate (ElSherief et al., 2021), contextual-abuse (Vidgen et al., 2021), tweet_offensive (Zampieri et al., 2019), 4chans (González-Pizarro and Zannettou, 2022), memes (Kiela et al., 2021); multi-class datasets for fine-grained hate speech classification: jigsaw (cjadams, 2017); latent_hatred (ElSherief et al., 2021), MMHS (Gomez et al., 2020).

Ideology & Stance Binary datasets for ideology classification: ibc (Gross et al., 2013); Ternary datasets for ideology & stance classification: vast (Allaway and McKeown, 2020); election_stance (Kawintiranon and Singh, 2021); media_ideology (Baly et al., 2020), SemEval (Mohammad et al., 2016), tweet_leg (Mou et al., 2021), tweet_cele (Wojcieszak et al., 2022).

Trustworthiness & Social Bias Binary datasets for trustworthiness classification: two-to-lie (Peskov et al., 2020); hypo-l (Zhang and Wan, 2022); neutralizing-bias-pairs (Pryzant et al., 2020).

Social Factors Binary datasets for social factors classification: Stanford Politeness (Fu et al., 2020), complaints (Preoţiuc-Pietro et al., 2019), empathy (Buechel et al., 2018), hayati_politeness (Hayati et al., 2021); Multi-class datasets for social factor classification: questionintimacy (Pei and Jurgens, 2020), pan (Pardo et al., 2018).

A.1.2 Raw Data Collection

We collect raw social media data with the help of previous related work (Kim et al., 2020). We then divide these raw data into the following datasets: hashtag_gen hashtag_choice, domain_explain, and personality_explain, each of which contains around 25k data. The ground truths of these datasets are generated by GPT-4V.

A.2 Instruction Construction

In this section, we will introduce the construction of instructional datasets for various tasks across modules. Specifically, we design a diverse array of prompts manually based on the collected dataset.

A.2.1 Knowledge & Comprehension Module

As discussed in §3.2, the Knowledge & Comprehension Module primarily encompasses classification tasks, for which we adapt different prompts to suit the various types of tasks.

Emotion There are two types of emotion classification tasks: coarse-grained emotion classification, which primarily involves determining whether a statement conveys a positive or negative sentiment, and fine-grained emotion classification, which entails identifying the presence of a specific emotion within a given statement.

Emotion Classification

Determine the emotion conveyed in the text following [Original Text], classifying it as either sadness, joy, love, anger, fear, or surprise.
[Original Text]: !<INPUT 0>!
Constraint: Provide a one-word answer.

Module	Category	Dataset	Size	Task Type	Data Type	Stage	Module	Category	Dataset	Size	Task Type	Data Type	Stage
Knowledge & Comprehension	Emotion	Css_Six_Emotion	30k	CLS	Text	SFT	Application	Emotion	Css_Six_Emotion_EXP	20k	GEN	Text	SFT
Knowledge & Comprehension	Emotion	Sentiment140	15k	CLS	Text	SFT	Application	Emotion	CARER_EXP	5K	GEN	Text	Eval
Knowledge & Comprehension	Emotion	CARER	5k	CLS	Text	Eval	Application	Humor	humor-pairs_EXP	15k	GEN	Text	SFT
Knowledge & Comprehension	Emotion	MVSA_Single	2.3k	CLS	Multi	SFT/Eval	Application	Humor	hahackathon#is_humor_EXP	6.1k	GEN	Text	Eval
Knowledge & Comprehension	Emotion	MVSA_Multiple	8.5k	CLS	Multi	SFT/Eval	Application	Hate Speech & Toxicity	jigsaw_EXP	25k	GEN	Text	SFT
Knowledge & Comprehension	Emotion	TumEmo	9.5k	CLS	Multi	SFT/Eval	Application	Hate Speech & Toxicity	tweet_offensive_EXP	4.6k	GEN	Text	SFT
Knowledge & Comprehension	Humor	reddit_jokes	4.1k	CLS	Text	SFT	Application	Hate Speech & Toxicity	contextual-abuse_EXP	1.9k	GEN	Text	Eval
Knowledge & Comprehension	Humor	puns	4k	CLS	Text	SFT	Application	Hate Speech & Toxicity	implicit-hate_EXP	8k	GEN	Text	Eval
Knowledge & Comprehension	Humor	short_jokes	9.9k	CLS	Text	SFT	Application	Hate Speech & Toxicity	latent_hatred_EXP	6.3k	GEN	Text	Eval
Knowledge & Comprehension	Humor	hahackathon#is_humor	8.3k	CLS	Text	Eval	Application	Ideology & Stance	ibc_EXP	4.3k	GEN	Text	SFT
Knowledge & Comprehension	Figurative Language	FLUTE	7.5k	CLS	Text	SFT	Application	Ideology & Stance	media_ideology_EXP	1k	GEN	Text	Eval
Knowledge & Comprehension	Figurative Language	sar	5k	CLS	Text	SFT	Application	Trustworthiness & Social Bias	neutralizing-bias-pairs_EXP	30k	GEN	Text	SFT
Knowledge & Comprehension	Figurative Language	tweet_irony	4.6k	CLS	Text	Eval	Application	Social Factors	domain_EXP	25k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Misinformation	climate_change	24k	CLS	Text	SFT	Application	Social Factors	personality_EXP	25k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Misinformation	cancer	0.6k	CLS	Text	Eval	Analysis	Figurative Language	sar_EXP	30k	GEN	Text	SFT
Knowledge & Comprehension	Misinformation	FakeNewsNet	6.5k	CLS	Multi	SFT/Eval	Analysis	Figurative Language	tweet_irony_EXP	2.2k	GEN	Text	Eval
Knowledge & Comprehension	Hate Speech & Toxicity	jigsaw	30k	CLS	Text	SFT	Analysis	Emotion	MVSA_Single_EXP	2.3k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Hate Speech & Toxicity	tweet_offensive	14k	CLS	Text	SFT	Analysis	Emotion	MVSA_Multiple_EXP	8.5k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Hate Speech & Toxicity	latent_hatred	6.3k	CLS	Text	Eval	Analysis	Emotion	TumEmo_EXP	9.5k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Hate Speech & Toxicity	4chans	2k	CLS	Multi	SFT/Eval	Analysis	Hate Speech & Toxicity	4chans_EXP	2k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Hate Speech & Toxicity	MMHS	7.5k	CLS	Multi	SFT/Eval	Analysis	Hate Speech & Toxicity	MMHS_EXP	7.5k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Hate Speech & Toxicity	hatefulmemes	4.3k	CLS	Multi	SFT/Eval	Analysis	Hate Speech & Toxicity	hatefulmemes_EXP	4.3k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Ideology & Stance	ibc	4.3k	CLS	Text	SFT	Analysis	Social Factors	PAN18_EXP	15k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Ideology & Stance	vast	18k	CLS	Text	SFT	Evaluation	Ideology & Stance	tweet_leg_EXP	1k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Ideology & Stance	election_stance	1.7k	CLS	Text	SFT	Evaluation	Ideology & Stance	tweet_cele_EXP	0.6k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Ideology & Stance	media_ideology	3.5k	CLS	Text	Eval	Evaluation	Misinformation	mrf_headlines_EXP	2k	GEN	Text	SFT
Knowledge & Comprehension	Ideology & Stance	tweet_leg	1k	CLS	Multi	SFT/Eval	Evaluation	Misinformation	FakeNewsNet_EXP	6.5k	GEN	Multi	SFT/Eval
Knowledge & Comprehension	Ideology & Stance	tweet_cele	0.6k	CLS	Multi	SFT/Eval	Evaluation	Trustworthiness & Social Bias	rumor_EXP	0.9k	GEN	Text	Eval
Knowledge & Comprehension	Trustworthiness & Social Bias	two-to-lie	11k	CLS	Text	SFT	Evaluation	Detoxifying Content	jigsaw_EXP	25k	GEN	Text	SFT
Knowledge & Comprehension	Trustworthiness & Social Bias	hypo-l	3.2k	CLS	Text	Eval	Evaluation	Detoxifying Content	contextual-abuse_EXP	1.9k	GEN	Text	Eval
Knowledge & Comprehension	Social Factors	Stanford Politeness	11k	CLS	Text	SFT	Evaluation	Detoxifying Content	implicit-hate_EXP	8k	GEN	Text	Eval
Knowledge & Comprehension	Social Factors	complaints	3.4k	CLS	Text	SFT	Evaluation	Depolarizing Language	ibc_EXP	4.3k	GEN	Text	SFT
Knowledge & Comprehension	Social Factors	empathy	1.8k	CLS	Text	SFT	Evaluation	Depolarizing Language	media_ideology_EXP	1k	GEN	Text	Eval
Knowledge & Comprehension	Social Factors	hayati_politeness	0.3k	CLS	Text	Eval	Creation	Invert Opinion	semeval_EXP	3k	GEN	Text	SFT
Knowledge & Comprehension	Social Factors	questionintimacy	2.2k	CLS	Text	Eval	Creation	Reverse ideology	ibc_EXP	4.3k	GEN	Text	SFT
Knowledge & Comprehension	Social Factors	PAN18	15k	CLS	Multi	SFT/Eval	Creation	Reverse ideology	media_ideology_EXP	1k	GEN	Text	Eval
Knowledge & Comprehension	Social Factors	hashtag_choice	25k	CLS	Multi	SFT/Eval	Creation	Social Factors	hashtag_gen_EXP	25k	GEN	Multi	SFT/Eval

Table 6: Composition of data for different modules

Multimodal Emotion Classification

This image is associated with the following caption: !<INPUT 0>!.
What sentiment does this combination convey? Positive, neutral, or negative? This is for research purposes.
CONSTRAINTS: only output one word from [positive, neutral, negative].

Humor The classification of humor is a binary classification task, which involves determining whether a given text is categorized as humor or not humor based on its content.

Humor Classification

Assess the provided [Original Text] to determine if it can be categorized as 'humor' or 'not humor'.
[Original Text]: !<INPUT 0>!
Constraint: Deliver a succinct evaluation, selecting either 'humor' or 'not humor'.

Figurative Language The classification task of figurative language is twofold: the first type is coarse classification, which determines whether the text contains figurative language, and the second type is fine classification, which identifies the specific type of figurative language used in the text.

Figurative Language Classification

Examine the text following [Original Text] for sarcasm. If the meaning contrasts with its literal interpretation, involves a situation of appearance versus reality, or carries a sarcastic tone, classify it as sarcasm. Otherwise, designate it as not-sarcasm.
[Original Text]: !<INPUT 0>!
Constraint: Provide a single-word response.

Figurative Language Classification

Analyze the [premise] to identify if the [hypothesis] represents sarcasm, creative paraphrase, metaphor, idiom, or simile.
[hypothesis]: !<INPUT 0>!
[premise]: !<INPUT 1>!
constraint: Single word answer

Misinformation The classification task of misinformation primarily involves identifying given news headlines or text-image pairs, determining whether they represent true information or false information.

Misinformation Classification

Appraise whether the given article [headline] is perceived as misinformation or real based on the [writer_intent] and [reader_action].
[writer_intent]: !<INPUT 0>!
[reader_action]: !<INPUT 1>!
[headline]: !<INPUT 2>!
Please answer with single word.

Multimodal Misinformation Classification

This news including an image is associated with the following caption:
!<INPUT 0>!
Question: Is this news true or false?
Choices: [true, false]

Hate Speech & Toxicity The classification task of Hate Speech & Toxicity is bifurcated into two categories: coarse classification, which determines whether a given text or text-image pair is offensive, and fine classification, which identifies the specific type of hate speech classification.

Hate Speech Classification

[Original text]: !<INPUT 0>!
Evaluate the given [Original Text] to ascertain whether its content should be classified as 'offensive' or 'not offensive'.
[Original Text]: !<INPUT 0>!
Restriction: Provide a concise response, choosing either 'offensive' or 'not offensive'.

Hate Speech Classification

[Original text]: !<INPUT 0>!
Identify the type of hate speech in the text following [original text], labeling it as either white-grievance, threatening, inferiority, stereotypical, incitement irony or other.
Restriction: Use only one word for your response.

Multimodal Hate Speech Classification

This image is associated with the following caption: '!<INPUT 0>!'.
Does this combination exhibit any elements of hate speech?
Choices: [true, false]

Multimodal Hate Speech Classification

This image is associated with the following caption: '!<INPUT 0>!'.
Does this combination exhibit any elements of hate speech? If so, which hate speech type does it belong to?
Choices: [NotHate, Racist, Sexist, Homophobe, Religion, OtherHate]

Ideology & Stance The classification task of Ideology & Stance primarily involves analyzing the ideological orientation of a given text or text-image pair, determining whether it aligns with liberal or conservative perspectives.

Ideology Classification

[Original text]: !<INPUT 0>!
Analyze the political orientation reflected in the provided text [Original Text] and categorize it as either "Liberal" or "Conservative".
[Original Text]: !<INPUT 0>!
Note: Provide a response using only one of the two specified categories: "Liberal" or "Conservative".

Multinodal Ideology Classification

This image is posted by a !<INPUT 0>! and is associated with the following caption: '!<INPUT 1>!'.
Question: What ideology does this !<INPUT 0>! belong to?
Choice: [left, center, right].

Trustworthiness & Social Bias The classification task of Trustworthiness & Social Bias primarily involves detecting the veracity of statements or determining whether they are exaggerated.

Trustworthiness Classification

Examine the given [Original Text] from an actual conversation to assess its truthfulness. Decide whether the statement is a 'truth' or a 'lie'.
[Original Text]: !<INPUT 0>!
Note: Please provide a brief response, choosing 'truth' or 'lie'.

Trustworthiness Classification

Evaluate [Original Text] to find hyperbole. If there are exaggerated statements, over-the-top expressions, or intentional exaggeration, mark it as Hyperbole. Otherwise, label it as Not-Hyperbole.
[Original Text]: !<INPUT 0>!

Social Factors The classification task of social factors encompasses a variety of task types, such as determining whether a given statement is polite, whether the statement demonstrates empathy or complaint, assessing the level of intimacy in a conversation, and the selection and generation of hashtags.

Social Factors Classification

Examine the [Original Text] for its overall tone, determining its classification as 'polite' or 'impolite'.
[Original Text]: !<INPUT 0>!
Instruction: Provide a straightforward response, selecting 'polite' or 'impolite'.

Social Factors Classification

Review the supplied [Original Text] to decide if it shows signs of 'empathy' or the absence thereof.
[Original Text]: !<INPUT 0>!
Obligation: Give a terse verdict, choosing between 'empathy' or 'not empathy'.

Social Factors Classification

Evaluate the given [Original Text] to ascertain whether it falls under the classification of 'complaint' or 'not complaint'.
[Original Text]: !<INPUT 0>!
Instruction: Provide a brief and clear decision, opting for either 'complaint' or 'not complaint' as the suitable categorization.

Social Factors Classification

Determine the intimacy level in the provided [Original Text]. Classify it as Very-intimate, Intimate, Somewhat-intimate, Not-very-intimate, Not-intimate, or Not-intimate-at-all using the following criteria.
criteria:
Very-intimate: the text involves a deeply personal or private matter, elicits a strong emotional response, or requires sharing sensitive information.
Intimate: the text involve sharing personal preferences, experiences, or opinions that go beyond surface-level topics.
Somewhat-intimate: the text touches on personal matters to some extent but is not as deep.
Not-very-intimate: the text discusses general or non-personal topics.
Not-intimate: the text is unrelated to personal matters or feelings.
Not-intimate-at-all: the text is entirely unrelated to personal matters and is more factual or transactional.
[Original Text]: !<INPUT 0>!
Constraint: Provide a single-word response.

Multimodal Social Factors Classification

This image and the following caption are from the same user: '!<INPUT 0>!'
Is the user likely to be male or female?
Pick your answer from [male, female].

Multimodal Social Factors Classification

This image is associated with the following caption by an Instagram user.
caption: !<INPUT 0>!
Which of the following hashtags BEST describes this post?
Choices: [!<INPUT 1>!, !<INPUT 2>!, !<INPUT 3>!, !<INPUT 4>!]
Constraints: only choose ONE hashtag from the Choice, and # should be included.

A.2.2 Application Module

As discussed in §3.3, the primary function of the Application Module is to interpret the ground truth labels of a given text.

Emotion The task within the "Application Module" related to emotions involves extracting the trigger that elicits a specific emotion, given the ground truth label of a provided text.

Emotion Trigger Extraction

The provided statement conveys a !<INPUT 1>! emotion. Kindly identify the stimuli that evoke this emotion.
[sentence]: !<INPUT 0>!

Humor The task within the "Application Module" related to humor is to provide corresponding explanations for statements labeled as humor in the ground truth data.

Humor Explanation

Consideration is given to the sentence being categorized as humor. Please elucidate the reasoning behind this classification.
[sentence]: !<INPUT 0>!

Hate Speech & Toxicity The task within the "Application Module" related to Hate Speech is aimed at providing explanations for texts classified as a certain type of Hate Speech.

Hate Speech Explanation

The sentences below are flagged for !<INPUT 1>! concerns. Please provide a concise explanation.
[sentence]: !<INPUT 0>!

Ideology & Stance The task within the "Application Module" regarding Ideology is to furnish corresponding explanations for texts categorized under a certain ideology (liberal or conservative).

Ideology Explanation

The following sentence suggests a perspective aligned with !<INPUT 1>!; Please provide a concise explanation.
[sentence]: !<INPUT 0>!

Trustworthiness & Social Bias The task of assessing trustworthiness and bias within the "Application Module" involves analyzing two given texts to determine which one exhibits greater bias.

Social Bias Explanation

Here we have two sentences. Kindly explain in a brief manner why !<INPUT 2>! is short.
[sentence]: !<INPUT 0>!
[sentence]: !<INPUT 1>!

Social Factors The social factor task within the application module consists of tasks to explain a user's domain or personality given a text-image pair post by the user.

Multimodal Social Factors Explanation

This image is linked with the following caption provided by a user.
Caption: !<INPUT 0>!
What is the user's professional field? Please explain in one sentence.

Multimodal Social Factors Explanation

This image is associated with the following caption by an Instagram user.
caption: !<INPUT 0>!
What's the personality of this user according to the post?
Constraints: First give the personality and explain it in one sentence.

A.2.3 Analysis Module

Figurative Language The task of Figurative Language in the Analysis Module involves enabling the model to analyze whether a text contains figurative language without the aid of known labels and to provide corresponding interpretations.

Figurative Language Analysis

Interpret the metaphorical or symbolic use of language in the following hypothesis in a single sentence.
[Hypothesis]: !<INPUT 0>!

Emotion The task of Emotion in the Analysis Module asks the model to generate the emotion or sentiment directly without any labels given.

Multimodal Emotion Analysis

This image is associated with the following caption: '!<INPUT 0>!'.
What fine-grained emotion does this combination convey?

Hate Speech & Toxicity The task of Hate Speech & Toxicity in the Analysis Module asks the model to identify whether the text-image pair contains any hate speech directly without any labels given.

Multimodal Hate Speech Analysis

This image is associated with the following caption: '!<INPUT 0>!'.
Does this combination exhibit any elements of hate speech? If so, which hate speech type does it belong to?

Social Factors The task of Social Factors in the Analysis Module asks the model to identify the gender of the user given the text-image pair without labels given.

Multimodal Social Factors Analysis

Determine the gender of the user given the following information.
This image and the following caption are from the same user: '!<INPUT 0>!'

A.2.4 Evaluation Module

Ideology & Stance The task of Stance & Ideology in the Evaluation Module asks the model to identify the stance of the user given the text-image pair without labels given.

Multimodal Ideolog & Stance Evaluation

This image is associated with the following caption: '!<INPUT 0>!'.
It is posted by a politician. What ideology does the politician belong to?

Misinformation The task of Misinformation within the Evaluation Module is aimed at interpreting the deep-seated implications of news headlines.

Misinformation Evaluation

Deduce the underlying implication of the news headline below. Provide a brief response, similar in style to 'some masks are better than others.'
[HEADLINE]: !<INPUT 0>!

Multimodal Misinformation Evaluation

This news is accompanied by an image and has the following description:
!<INPUT 0>!

Can the authenticity of this article be verified?

Trustworthiness & Social Bias The task of Trustworthiness within the Evaluation Module aims to detect rumors and provide corresponding explanations.

Trustworthiness Evaluation

The sentence implies that it embodies the quality of being !<INPUT 1>!. Kindly offer an explanation for the basis of this statement.
[sentence]: !<INPUT 0>!

Detoxifying Content The task of "Detoxifying Content" within the Evaluation Module aims to rewrite hate speech, reducing its toxicity.

Detoxifying Content

Revise this sentence to make it more positive while keeping its original intent as much as possible.
[sentence]: !<INPUT 0>!

Depolarizing Language The task of Depolarizing Language in the Evaluation Module is aimed at depolarizing ideological discourse.

Depolarizing Language

Rewording the following sentence to make it less polarizing.
[sentence]: !<INPUT 0>!

A.2.5 Creation Module

Reverse Ideology The task of Reverse Ideology in the Creation Module involves providing the model with a text characterized by a specific ideology (either liberal or conservative) and prompting the model to produce statements on the same topic that reflect the opposite ideology.

Depolarizing Language

Reframe this !<INPUT 1>! speech from a !<INPUT 2>! perspective, ensuring the core theme remains the same.
[sentence]: !<INPUT 0>!

Social Factors The task of Social Factors in the Creation Module involves providing the model with a text-image pair and prompting the model to generate three hashtags that best summarize the post.

Multimodel Hashtag Generation

This image is associated with the following caption by an Instagram user.
Caption: !<INPUT 0>!
Try to generate no more than 3 hashtags that best fit this post.
Constraints: the hashtags should begin with #.
Output Format: #hashtag_1, # hashtag_2, # hashtag_3

B Training Details

B.1 Computational resources

All of our experiments were conducted on an Ubuntu 22.04.3 machine installed with NVIDIA RTX 3090 and 4090 GPUs. The Python packages used in our experiments include Pytorch 2.1.1, Transformers 4.33.0, and deepspeed 0.11.1.

B.2 Details on large language model instruction tuning

As mentioned in §4.4, we employ the QLoRA method (Dettmers et al., 2023) with FastChat (Zheng et al., 2023) for language model tuning. The specific settings for the hyper-parameters are presented in Table 7.

Hyper-parameters	Value
lora_r	128
lora_alpha	256
per_device_train_batch_size	8
gradient_accumulation_steps	2
learning_rate	2e-5
weight_decay	0.
warmup_ratio	0.05
lr_scheduler_type	cosine
tf32	True
model_max_length	2048
q_lora	True
flash_attn	True

Table 7: Hyper-parameters of Language Model Tuning

B.3 Details on Q-former instruction tuning

As mentioned in §4.4, we tuned our connection module following the pipeline of LAVIS (Li et al., 2023a). The specific settings for the hyperparameters are presented in Table 8.

Hyper-parameters	Value
init_lr	3e-5
min_lr	1e-5
lr_sched	linear_warmup_cosine_lr
weight_decay	0.02
max_epoch	3
batch_size_train	1
batch_size_eval	1
num_workers	1
freeze_vit	True

Table 8: Hyperparameters of Connection Module Tuning.

C Experiment Results on Each Dataset

C.1 Textual Datasets

Experiment results on each dataset in textual tasks are shown in Table 9 and Table 10.

C.2 Multimodal Datasets

Experiment results on each dataset in multimodal tasks are shown in Table 11 and Table 12.

Datasets	SoMeLVLM Accuracy	Vicuna Accuracy	Llama2 Accuracy	Chatglm2 Accuracy
Twitter_emotion	**80.66**	35.86	40.54	41.20
hahackathon#is_humor	60.47	41.08	**61.31**	36.94
tweet_irony	**61.70**	47.08	53.77	52.05
misinfo_cancer	**70.38**	59.23	41.11	47.21
latent_hatred	**22.20**	11.94	12.84	14.67
media_ideology	**45.23**	34.15	37.77	30.08
hypo-l	43.52	36.60	59.21	**68.44**
hayati_politeness	**89.68**	70.63	49.69	83.43
question intimacy	**21.09**	14.73	13.53	13.03

Table 9: Classification results on each dataset in the textual experiment.

	SoMeLVLM			Vicuna			Llama2			Chatglm2		
Dataset	BLEU	ROUGE	Score	BLEU	ROUGE	Score	BLEU	ROUGE	Score	BLEU	ROUGE	Score
twitter_emotion_EXP	**26.96**	**51.88**	**3.63**	7.97	31.31	3.23	4.25	23.50	2.99	6.60	29.47	3.05
hahackathon#is_humor_EXP	**13.81**	**42.84**	**3.38**	10.49	36.21	3.24	6.36	28.37	2.48	8.98	34.49	2.37
tweet_irony_EXP	**23.77**	**45.42**	**3.02**	8.03	31.55	2.57	10.39	31.32	2.73	7.20	29.07	2.06
contextual-abuse#IdentityDirectedAbuse_EXP	**18.10**	**43.36**	**3.55**	6.49	30.80	3.46	1.69	17.72	1.96	4.24	27.19	2.60
contextual-abuse#PersonDirectedAbuse_EXP	**18.56**	**45.38**	**3.72**	6.86	30.22	3.62	1.38	15.28	1.55	4.50	27.53	2.71
implicit-hate#explicit_hate_EXP	**20.76**	**47.49**	**3.85**	8.09	33.11	3.83	2.11	19.02	2.09	4.77	28.90	3.42
implicit-hate#implicit_hate_EXP	**14.87**	**39.78**	3.52	6.82	31.37	**3.61**	1.78	17.43	1.97	4.23	28.33	2.94
latent_hatred_EXP	**13.89**	**39.51**	3.58	6.08	30.72	**3.62**	1.99	17.60	2.13	4.75	28.29	3.02
media_ideology_EXP	**14.60**	**39.49**	**3.43**	9.36	32.78	3.41	4.75	25.01	2.78	6.59	29.94	2.86
rumor#rumor_bool_EXP	**12.37**	**39.06**	**3.59**	9.70	34.13	3.13	4.73	26.54	2.82	9.25	34.35	2.73
contextual-abuse#IdentityDirectedAbuse_EXP	**28.11**	**48.68**	**3.00**	11.00	28.47	2.60	1.57	11.54	1.23	6.50	22.85	2.00
contextual-abuse#PersonDirectedAbuse_EXP	**29.64**	**49.39**	**3.08**	11.37	28.21	2.66	1.67	12.13	1.34	6.62	23.25	2.08
implicit-hate#explicit_hate_EXP	**22.98**	**43.78**	**2.50**	7.15	23.76	2.07	0.80	9.24	0.90	5.92	22.63	1.74
implicit-hate#implicit_hate_EXP	**27.77**	**49.18**	**2.97**	12.21	31.38	2.69	1.21	10.85	1.07	8.30	26.94	2.18
media_ideology_EXP	**23.54**	**45.47**	**3.28**	22.31	42.72	3.26	8.40	26.72	2.21	13.33	35.66	2.80
media_ideology_EXP	**44.09**	**61.96**	**3.41**	33.40	51.76	2.981	20.54	38.06	2.04	21.91	42.27	2.80

Table 10: Generation results on each dataset in the textual experiment.

	SoMeLVLM		Instructblip$_V$		Instructblip$_F$		Blip2		Llava		Minigpt4	
Datasets	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc	Acc*	Acc
4chans	75.00	75.00	55.49	50.50	57.47	56.75	56.00	56.00	79.49	15.50	66.14	41.50
MMHS	67.40	67.40	22.01	13.60	31.65	31.40	34.00	34.00	29.53	11.40	18.08	9.40
FakeNewsNet	82.60	82.60	47.55	13.60	80.78	79.00	80.60	80.60	84.67	25.40	65.30	54.20
hatefulmemes	75.80	75.80	50.13	39.60	63.50	58.80	67.20	67.20	56.25	3.60	55.33	21.80
MVSA_single	76.05	76.05	58.27	53.88	70.09	69.62	70.07	70.07	62.50	4.43	57.39	29.27
MVSA_multiple	67.60	67.60	59.28	55.60	65.12	64.60	64.40	64.40	65.21	3.00	62.31	33.40
PAN	69.00	69.00	68.92	55.00	64.92	64.40	64.80	64.80	54.37	11.20	56.71	41.40
TumEmo	48.19	48.10	46.50	37.80	42.70	40.45	40.04	40.04	33.43	22.36	40.19	25.81
tweet_leg	83.45	64.36	65.25	48.94	62.05	54.79	55.32	55.32	66.67	2.12	50.00	9.04
tweet_cele	58.24	41.41	37.84	32.81	41.41	32.03	50.78	50.78	25.00	0.78	30.56	8.59
hashtag_choice	99.38	65.64	91.30	26.64	98.00	82.88	99.13	97.25	90.91	2.11	71.57	30.87

Table 11: Classification results on each dataset in the multimodal experiment.

	SoMeLVLM			Instructblip$_V$			Instructblip$_F$			Blip2			Llava			Minigpt4		
Datasets	BLEU	ROUGE	GPT	BLEU	ROUGE	GPT	BLEU	ROUGE	GPT	BLEU	ROUGE	GPT	BLEU	ROUGE	GPT	BLEU	ROUGE	GPT
4chans_EXP	27.42	49.76	3.33	0.74	3.34	1.60	0.42	4.23	1.51	1.29	5.18	1.63	0.46	6.06	1.27	0.54	9.91	3.15
hatefulmemes_EXP	33.37	48.60	2.83	0.53	3.17	2.37	0.23	3.39	2.63	0.15	1.10	2.13	0.39	5.07	1.29	0.36	9.19	1.95
MMHS_EXP	32.34	40.68	3.49	0.69	2.87	1.47	0.07	0.75	2.07	0.41	0.46	1.76	0.22	2.43	1.14	0.38	7.41	1.90
FakeNewsNet_EXP	24.06	43.22	2.94	1.09	6.21	2.84	0.05	0.81	2.85	0.02	1.89	2.72	0.00	0.01	0.81	0.69	12.15	2.18
PAN_EXP	35.42	61.05	3.48	0.39	6.21	1.00	1.17	22.16	2.88	0.15	21.39	3.17	1.47	9.81	1.54	0.42	23.95	1.64
hashtag_gen	2.94	8.51	1.10	0.95	1.07	0.80	0.60	1.78	1.14	1.52	0.53	1.12	1.96	2.43	1.08	0.85	4.97	1.06
domain_explain	10.25	31.94	3.35	0.57	13.27	1.67	1.29	15.80	2.09	0.92	13.98	1.71	1.77	19.35	2.03	1.78	20.57	1.83
personality_explain	9.33	29.98	3.50	1.62	15.52	2.40	1.56	18.65	2.34	0.45	12.06	1.53	2.35	19.62	2.54	1.73	19.30	1.85
MVSA_multiple_EXP	42.91	60.58	3.80	1.15	9.64	2.24	0.23	19.26	3.65	0.22	22.74	3.82	0.88	6.73	1.61	0.71	11.63	2.79
MVSA_single_EXP	39.38	59.12	3.78	0.85	6.60	1.88	0.23	17.31	3.36	0.21	21.43	3.59	0.83	6.53	1.51	0.68	11.87	2.55
TumEmo_EXP	30.66	41.92	3.03	0.56	5.54	1.75	0.39	4.49	2.09	0.06	0.28	1.88	0.21	3.95	0.64	0.26	8.93	1.79
tweet_cele_EXP	19.02	37.45	2.75	0.41	3.53	1.14	0.86	8.06	1.07	0.24	2.78	2.23	0.76	6.40	0.54	0.29	13.26	0.59
tweet_leg_EXP	29.14	44.62	3.82	0.79	6.24	1.93	0.69	8.65	1.99	0.26	5.92	2.42	1.44	11.06	1.66	0.34	12.10	1.75
domain_ood	10.41	31.85	3.38	0.49	11.73	1.62	1.26	15.11	2.04	0.88	13.85	1.66	2.07	20.23	1.97	1.89	20.88	1.74
personality_ood	9.95	30.20	3.52	1.79	16.33	2.53	1.75	18.70	2.29	0.41	11.89	1.56	2.51	19.97	2.58	2.07	20.57	1.95

Table 12: Generation results on each dataset in the multimodal experiment.