

Supplementary Materials of WisdoM

Wenbin Wang*
School of Computer Science
Wuhan University
Wuhan, China
wangwenbin97@whu.edu.cn

Yong Luo†
Wuhan University &
Hubei LuoJia Laboratory
Wuhan, China
luoyong@whu.edu.cn

Liang Ding*
The University of Sydney
Sydney, Australia
liangding.liam@gmail.com

Han Hu
Beijing Institute of Technology
Beijing, China
hhu@bit.edu.cn

Li Shen
Sun Yat-Sen University
Shenzhen, China
mathshenli@gmail.com

Dacheng Tao
Nanyang Technological University
Singapore, Singapore
dacheng.tao@ntu.edu.sg

CCS Concepts

• **Information systems** → *Sentiment analysis*.

ACM Reference Format:

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Supplementary Materials of WisdoM. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 8 pages. <https://doi.org/3664647.3681403>

A Experiment Details

A.1 Details of Datasets

In this work, we conduct experiments covering aspect-level (Twitter2015 and Twitter2017) and sentence-level (MSED) MSA benchmarks. Additionally, we manually selected multimodal tweets that mention politicians from the Twitter2015 and Twitter2017 datasets to construct the Politician Twitter dataset [16]. We present the statistics of all datasets in Table 1.

A.2 Model Details

To verify the effectiveness of our WisdoM, we apply it within two standard frameworks for modeling the MSA tasks. For task-specific framework, we conduct experiments using the AoM (105M) [19] and ALMT (112.5M) [17]. For general-purpose framework, we experiment on Qwen-VL (9.6B) [1], LLaVA-v1.5 (13B) [6] and MMICL (14B) [18]. The detailed model information is listed in Table 2.

*Equal contribution.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/3664647.3681403>

A.3 Training Details

For a fair comparison, we fine-tune models introduced in Section 4.1.2 on target datasets. We optimize our model with AdamW [7]. The learning rate is grid searched in {1e-5, 2e-5, 7.5e-5, 1e-4}, and batch size is in {16, 32, 256}. It should be noted that *contexts* are not incorporated during the training stage.

A.4 Prompt Templates of World Knowledge

In Table 3, we list the prompt templates which are provided by ChatGPT in prompt templates generation stage. In the context generation stage, we replace “[x]” with sentence s_i and a special image token (e.g., “<image>”) is inserted at the beginning of the prompt template.

A.5 Pseudo Code for Contextual Fusion

Algorithm 1 provides the pseudo-code of Contextual Fusion. The simplicity of our method requires only a few lines of code. Please note that in the specific implementation, we use the formula:

$$\delta_i = 2 \times P_i^{max} + P_i^{min} - 1, \quad (1)$$

where Eq. (1) is equivalent to $P_i^{max} - P_i^{sec}$. Additionally, the implementation of this equation is more efficient.

Algorithm 1: Python-style pseudo-code for Contextual Fusion.

```
# p_o: numpy array, represents  $P_i$ 
# p_c: numpy array, represents  $\hat{P}_i$ 
# alpha: float, the threshold of choosing hard sample
# beta: float, interpolation coefficient
def contextual_fusion(p_o, p_c, alpha, beta):
    # step1: calculate delta
    delta = 2 * np.max(p_o) + np.min(p_o) - 1 #  $P_i^{max} - P_i^{sec}$ 
    if delta > alpha:
        return p_o
    # step2: calculate the final prediction
    p_f = p_o + beta * (p_c - p_o)
    return p_f
```

A.6 Details of Hand-Crafted and LLM-Generated Prompt Templates

To compare hand-crafted and LLM-generated prompt templates, we employ human annotators and ChatGPT to design three hand-crafted prompt templates respectively, focusing on historical world knowledge. We list prompt templates generated by hand-crafted and LLM-generated in Table 4 and Table 5.

Table 1: Dataset statistics.

	Twitter2015			Twitter2017			Politician Twitter			MSED		
	#Train	#Dev	#Test	#Train	#Dev	#Test	#Train	#Dev	#Test	#Train	#Dev	#Test
<i>Negative</i>	368	149	113	416	144	168	874	161	323	1939	308	613
<i>Neutral</i>	1883	679	607	1638	517	573	4506	933	449	1664	294	569
<i>Positive</i>	928	303	317	1508	515	493	3221	618	225	2524	419	860
<i>Total</i>	3179	1122	1037	3562	1176	1234	8601	1712	997	6127	1021	2042

Table 2: Information of all models used in this study.

Model	Model Type	Source
AoM	Task-specific	https://github.com/SilyRab/AoM
ALMT	Task-specific	https://github.com/Haoyu-ha/ALMT
LLaVA-v1.5	General-purpose	https://huggingface.co/liuhaotian/llava-v1.5-13b
Qwen-VL	General-purpose	https://github.com/QwenLM/Qwen-VL
MMICL	General-purpose	https://huggingface.co/BleachNick/MMICL-Instructblip-T5-xxl

Table 3: Example of the prompt template generated in stage 1 and used in stage 2.

Type name	Prompt template
Biographical	Delve into the backgrounds of individuals associated with the image and text. Explore the biographies of artists, authors, or other relevant figures, and discuss how their life experiences shaped the creation and interpretation of the work. Sentence: [x]
Cultural	Explore how the image and sentence reflect or represent aspects of a particular culture. Discuss the cultural significance, traditions, or values implied by the elements in the image and sentence. Sentence: [x]
Environmental	Examine the environmental elements within the image and sentence, discussing ecological factors, environmental changes, or the relationship between human activities and the depicted setting. Sentence: [x]
Historical	Give you an image and sentence, you can provide historical context, important events, and relevant background information related to the image and sentence. Sentence: [x]
Political	Examine the politics during the time the image and text were created. Discuss any political events, movements, or ideologies that may have influenced the content and tone of the work. Sentence: [x]
Scientific	Investigate the scientific elements within the image, delving into discoveries, advancements, or breakthroughs related to the subject matter mentioned in the sentence. Sentence: [x]
Social	Investigate the image and text as a form of social commentary. Analyze how the work reflects or critiques social issues, norms, or inequalities prevalent at the time of creation. Sentence: [x]
Financial	Give you a sentence and image, you should provide related financial knowledge. Sentence: [x]

A.7 RAG Experimental Setup

In Section 4.4.3, we compare our WisdoM with two RAG-based methods: (1) a naive RAG [5], which initially searches for relevant documents related to a given question and then employs a generator to predict an answer; (2) PKG [8], an advanced RAG method that incorporates a knowledge-guided module, allowing for information retrieval without modifying the parameters of language models. The experimental setting is described below.

Knowledge Sources. The source of knowledge for our experiments is the Wikipedia-Image-Text (WIT) dataset [13], which has published in 2021. This dataset comprises images from Wikipedia, along with their alt-text captions and contextualized text passages.

Methods. For naive RAG, we use Contriever-MSMARCO [4] as the textual retriever and CLIP-ViT [11] as the visual retriever. Specifically, we utilize CLIP-ViT and Contriever-MSMARCO for encoding

query q and knowledge source, and employ Maximum Inner Product Search (MIPS) [3] to find the five nearest neighbors (*knowledge*) to q within the knowledge source. For PKG, we use LLaVA-v1.5 (13B) as knowledge-guided module fine-tuning on WIT and then generate the *knowledge* according to the image v and sentence s . Subsequently, we directly predict sentiment polarity by incorporating *knowledge*.

Evaluation Setting. To evaluate the relevance of the context to a specific image v and sentence s , we employ LLM-based metric, *i.e.*, LLM-as-a-Judge [2]. Specifically, a prompt is crafted for GPT-4V [10] to assess the winning rates of our context in comparison to those derived from RAG-based methods. The detailed prompt can be found in Table 6. We also evaluate performance on MSA tasks.

Table 4: Hand-Crafted prompt templates. For each human annotators, we provide an example of image-sentence pair with meta-prompt to better design the prompt templates.

Hand-Crafted	
Meta-prompt	Please provide prompt template that large vision-language model can generate the historical knowledge based on the image and sentence.
Human-Crafted	<ol style="list-style-type: none"> 1. Please give historical information about the image and sentence. 2. Can you provide historical context related to the image and sentence? 3. Please analyze the significance depicted in the image and text, and discuss how they embody and convey historical knowledge.

Table 5: LLM-Generated prompt templates. The meta-prompt is used by ChatGPT to generate prompt templates.

LLM-Generated	
Meta-prompt	Please provide prompt template that large vision-language model can generate the historical knowledge based on the image and sentence.
LLM-Generated	<ol style="list-style-type: none"> 1. Based on the image and text provided, please generate historical knowledge. This can include background information, important events, cultural context, or any relevant historical details related to the image and sentence combination. 2. Provide historical knowledge related to the image and sentence, including important events, cultural significance, and background information that contextualizes the scene depicted. 3. Give you an image and sentence, you can provide historical context, important events, and relevant background information related to the image and sentence.

A.8 Calculation of $S_{c \rightarrow p}$ and $S_{i \rightarrow p}$

To measure the significance of *context* and original input (*i.e.*, image and sentence), we use $S_{c \rightarrow p}$ and $S_{i \rightarrow p}$ for highlighting critical token interactions. Following previous work [12, 14], we use the Taylor expansion [9] to calculate the score for each element of the attention matrix:

$$I_l = \sum_h |A_{h,l}^\top \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}}|, \quad (2)$$

where $A_{h,l}$ represents the value of the attention matrix of the h -th attention head in the l -th layer. The x represents the input, and $\mathcal{L}(x)$ is the loss function. We calculate the saliency matrix I_l for the l -th layer by averaging across all attention heads with $I_l(k, j)$ indicating the importance of the information flow from the j -th word to the

Table 6: The Evaluation Prompt we used for GPT-4V. [s] represents the input sentence. [x1] and [x2] represent the context generated by different methods.

Evaluation Prompt
<p>**System**: In this task, you will be asked to compare the relevance of two paragraphs to determine which one is more pertinent to the provided source sentence and image and benefits the sentiment analysis task the most. There are three options for you to choose from:</p> <ol style="list-style-type: none"> 1. Context1 is better. If you think Context 1 is more relevant to the source sentence and image and benefits the sentiment analysis task. 2. Context2 is better. If you think Context 2 is more relevant to the source sentence and image and benefits the sentiment analysis task. 3. Context1, Context2 are the same: If you think Context1, Context2 have the same relevance to the source sentence and image, then choose this option. <p>**Your answer is a JSON DICT that has one key: answer. For example: {"answer": "x. Context x is better."}**</p> <p>**INPUT**</p> <p>Source Sentence: “[s]”</p> <p>Context1: “[x1]”</p> <p>Context2: “[x2]”</p> <p>**OUTPUT**</p>

k -th word [14]. We propose two quantitative metrics based on I_l . The definitions of the two quantitative metrics are below.

$S_{c \rightarrow p}$, the mean significance of information flow from *context*(c) to the prediction label (p).

$$S_{c \rightarrow p} = \frac{\sum_{(k,j) \in C_{cp}} I_l(k, j)}{|C_{cp}|}, \quad (3)$$

$$C_{c,p} = \{(c, p) : c \in \text{context}\}.$$

$S_{i \rightarrow p}$, the mean significance of information flow from *image* (v) and *sentence* (t) to the prediction label (p).

$$S_{i \rightarrow p} = \frac{\sum_{(k,j) \in C_{ip}} I_l(k, j)}{|C_{ip}|}, \quad (4)$$

$$C_{i,p} = \{(i, p) : i \in [v, t]\}.$$

$S_{c \rightarrow p}$ and $S_{i \rightarrow p}$ indicate the intensity of information aggregation onto the prediction label. A high S demonstrates strong information for final decision-making.

B Additional Experiments

B.1 Training with Context

To investigate the impact of context incorporation during training, we conduct experiments with two different setups: 1) training with context, which involves the incorporation of context into training phase; 2) our WisdoM, to leverage world knowledge context during the inference phase. The results, depicted in Table 7, demonstrate that our WisdoM consistently outperforms the baseline and training with context across all models. Additionally, comparison with baseline, the performance of training with context experienced a

significant decline. The possible reason is that the noisy nature of the context used during training. We conjecture that developing a method to effectively filter out this noise could potentially ameliorate the performance of models trained with context.

Table 7: Comparison of WisdoM and training with context on Twitter2015 and Twitter2017. We utilize the historical world knowledge generated by LLaVA-v1.5 for training and observed significant decreases in AoM and LLaVA-v1.5.

Method	Twitter2015		Twitter2017	
	Acc.	Mac-F1	Acc.	Mac-F1
AoM				
Baseline	80.0	75.2	75.9	74.5
Training w/ Context	78.0	74.0	73.7	71.6
-w/ WisdoM	81.5	78.1	77.6	76.8
LLaVA-v1.5				
Baseline	77.9	74.3	74.6	74.3
Training w/ Context	78.4	74.4	73.3	73.5
-w/ WisdoM	78.9	75.6	75.6	75.3

B.2 Efficiency vs. Performance

To address concerns about efficiency, we apply WisdoM to Qwen-VL and measure the average inference time (second) on Twitter2015 and MSSED datasets. We use LMDeploy to speed up inference in stage 2. As shown, although the average inference time increased by 1.1 seconds, we gain an average increase of 1.9% F1 score, which is a fruitful tradeoff.

Table 8: The relationship between inference duration and F1.

	Twitter2015		MSSED	
	Duration (s)	Mac-F1	Duration (s)	Mac-F1
Qwen-VL 9B	0.2	70.0	0.2	89.4
-w/ WisdoM	1.8	71.3	0.8	91.8

B.3 Evaluating LVLMS with Prompt to Complete MSA Tasks

To explore the impact of different prompt strategies on LVLMS in MSA tasks, we conduct experiments using LLaVA-v1.5 and InternVL-v1.5 on Twitter2015 and MSSED. We compare the prompts used in Table 1 (in our paper), Chain-of-Thought (CoT) [15], and our WisdoM. As shown in Table 9, the results show that 1) CoT improves LVLMS performance on MSA tasks, and 2) WisdoM enhances performance when applied to SOTA LVLMS.

C Case Study

C.1 Example of Contexts

Examples of context from RAG-based methods and our WisdoM are presented in Table 10. In case 1, the *context* provided by RAG related to the Harry Potter and Forbidden Journey ride is generic,

Table 9: Comparison different prompt-based strategies on LVLMS in MSA tasks.

Method	Twitter2015		MSSED	
	Acc.	Mac-F1	Recall	Mac-F1
LLaVA-v1.5 13B	41.4	39.3	65.0	61.5
-w/ CoT	43.0 (+1.6)	42.8 (+3.5)	64.6 (-0.4)	61.0 (-0.5)
-w/ WisdoM	45.6 (+4.2)	44.9 (+5.6)	65.1 (+0.1)	61.6 (+0.1)
InternVL-v1.5 26B	54.4	55.2	61.6	57.1
-w/ CoT	55.2 (+0.8)	56.7 (+1.5)	61.9 (+0.3)	57.7 (+0.6)
-w/ WisdoM	56.5 (+2.1)	57.6 (+2.4)	61.7 (+0.1)	57.2 (+0.1)

mentioning it as a motion-based dark ride located within various Harry Potter-themed areas across Universal Studios parks worldwide. However, our WisdoM provides a more detailed description, which aligns more closely with the related sentence about taking a selfie on the way to Harry Potter Studios. It is evident that **our WisdoM offers more precise context, providing detailed descriptions of image elements. In contrast, RAG-based methods exhibit a lack of specificity in image details and demonstrate weak relevance to the associated sentences.**

C.2 Qualitative Examples of Aspect-Level MSA

We present two qualitative examples from Table 11 showcasing historical knowledge. Table 11 shows that: 1) **SOTA model primarily assess the visible elements in an image and sentence, failing to consider deeper historical knowledge; they only consider superficial information.** For instance, without additional *context*, AoM predicts the sentiment polarity of Aleppo as neutral in a given sample. However, when considering the historical context of Aleppo’s devastation during the war, which contrasts sharply with its bustling period before the war, the sentiment is accurately identified as negative. 2) **Irrelevant information (noise) in the context leads to errors in sentiment analysis, whereas our Contextual Fusion can effectively mitigate these effects.** For example, a tweet about Donald Trump is initially assessed as positive. However, this prediction changes with the inclusion of context: **“This cartoon is likely a political satire or a commentary...”**. With the aid of Contextual Fusion, LLaVA-v1.5 confidently assesses the sentiment as positive, allowing it to overlook the effect of *context*.

C.3 Qualitative Examples of Sentence-Level MSA

Table 12 illustrates two instances where incorporating scientific world knowledge enhances the MSSED dataset. In the first example, ALMT initially predicts negative, but incorporating information about the positive role of masks in epidemic prevention shifts the prediction to positive. In the second example, an excessive focus on the benefits of coffee leads to misclassification by LLaVA-v1.5. However, with the aid of the Contextual Fusion, LLaVA-v1.5 ultimately makes the correct prediction. We find that **incorporating relevant world knowledge not only enhances the precision of sentiment analysis, but may also introduce some noise. Nevertheless, our Contextual Fusion module efficiently mitigates the effects of noise.**

Table 10: Examples of contexts generated by RAG, PKG and our WisdoM. Our contexts provide a detailed description that matches the elements present in images.



Sentence	RT @ shanilpanara : Bus Selfie on the way to Harry Potter Studios @ WFCTrust @ NCSEast # ShareYourSummer	
Image		
RAG	Harry Potter and the Forbidden Journey is a motion-based dark ride located in The Wizarding World of Harry Potter-themed areas of Islands of Adventure in Orlando, Florida; Universal Studios Hollywood in Universal City, California, Universal Studios Japan in Osaka, Japan and is upcoming to Universal Studios Beijing in Beijing, China. The ride takes guests through scenes and environments in and around Hogwarts Castle from the Harry Potter series of books and films. Mark Woodbury, president of Universal Creative, described the ride as an in-depth look at the world of Harry Potter, which utilizes never-before-seen technology which transforms “the theme park experience as you know it”. The ride first opened at Islands of Adventure with The Wizarding World of Harry Potter on June 18, 2010; subsequent versions opened at Universal Studios Japan on July 15, 2014, and at Universal Studios Hollywood on April 7, 2016.	
PKG	A picture of a person with a camera and a big lens in front of them, taken by a tourist.	
WisdoM	The image shows a large group of people, including men and women, sitting on a bus. They are all wearing black and white clothing, which suggests that the photo might have been taken during a specific time period or event when black and white clothing was more common. The group appears to be diverse, with people of different ages and ethnicities, which indicates that the bus might be a public transportation vehicle or a special event bus. In the context of historical events, the black and white clothing could imply that the photo was taken during the early to mid-20th century when colour photography was not widely available. The group of people might be attending a social event, a gathering, or a public transportation service during that time. The presence of a diverse group of people on the bus also reflects the social dynamics and cultural diversity of the period.	
Sentence	RT @ wemssecnirp : Now I really am like Cleopatra # hypnotizing # findom # goddess	
Image		
RAG	Cleopatra has been the subject of literature, films, plays, television programs, and art. Only those with Wikipedia articles are cited.	
PKG	A portrait of Cleopatra, the Queen of Egypt, wearing a golden headdress and a necklace. The portrait was drawn in the style of Raphael and engraved on copper.	
WisdoM	The image shows a woman wearing a gold headband with a large, ornate design featuring eyes. This headband is likely inspired by ancient Egyptian or African culture, as it resembles the traditional headpieces worn by women in these regions. The eyes on the headband may symbolize protection, wisdom, or the ability to see beyond the physical realm. In ancient Egypt, headbands were often worn by both men and women as a symbol of status, power, and religious beliefs. They were made from various materials, such as gold, silver, or precious stones, and often featured intricate designs and symbols. Similarly, in African cultures, headbands have been used as a form of adornment and expression of identity, with designs and patterns often representing specific tribes, clans, or regions. The woman in the image is wearing a modern interpretation of these traditional headbands, which demonstrates the enduring influence of ancient cultures on contemporary fashion and design.	

Table 11: Qualitative examples of aspect-level MSA. Two cases with predictions by SOTA methods and our method on Twitter2015 and Twitter2017. We initially integrate only the *context* and then introduce Contextual Fusion. Clues to answering the input are highlighted in **green within the *Context*. The *Context* that will disrupt MSA is highlighted in **red**.**


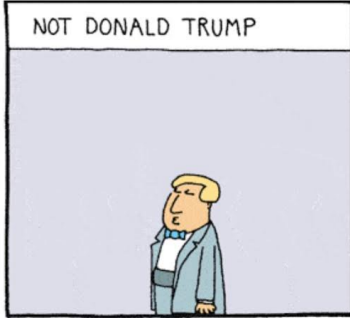


Sentence	RT @ AHedengren : # Aleppo before and after . # Syria .	
Image		
Aspect	Aleppo	
Context	<p>The image shows a comparison of the same street in Aleppo, Syria, before and after the devastating war. The before picture depicts a bustling and lively street with people walking and cars driving, while the after picture displays a desolate and destroyed street with rubble and ruins. The stark contrast between the two images highlights the devastating impact of the war on the city and its residents. Aleppo, once a thriving and historic city, has been severely affected by the ongoing Syrian Civil War. The conflict has led to the destruction of infrastructure, displacement of residents, and loss of life. The image serves as a powerful reminder of the human cost of war and the need for peaceful resolution to conflicts.</p>	
Label	Negative	
AoM	Neutral	
Is hard sample?	✓	
+ Context	Negative	
+ Contextual Fusion	Negative	
Sentence	Donald Trump will win in a landslide. * The cartoonist behind ‘ Dilbert ’ explains why.	
Image		
Aspect	Donald Trump	
Context	<p>The image features a cartoon of a man in a suit and tie, standing in front of a wall with a sign that says "Not Donald Trump." This cartoon is likely a political satire or a commentary on the 2016 United States presidential election, as Donald Trump was one of the main candidates running for office at that time. The cartoon could be a representation of the political climate, opinions, or reactions towards Donald Trump and his campaign. It might also be a way to express the artist's or the creator's perspective on the election, the political landscape, or the candidates involved.</p>	
Label	Positive	
LLaVA-v1.5	Positive	
Is hard sample?	✗	
+ Context	Negative	
+ Contextual Fusion	Positive	

Table 12: Qualitative examples of MSED. We initially integrate only the *context* and then introduce Contextual Fusion. Clues to answering the input are highlighted in **green within the *Context*. **Red** represents incorrect prediction.**

Sentence	Group of teenagers posing showing their protective face masks during COVID-19 coronavirus epidemic spread.
Image	
Context	The image features a group of teenagers posing with their protective face masks, which are an essential part of personal protective equipment (PPE) during the COVID-19 pandemic. The use of face masks has been a crucial measure to reduce the transmission of the virus, as they help prevent respiratory droplets from spreading into the air. Scientifically, the effectiveness of face masks in preventing the spread of COVID-19 has been extensively studied. Research has shown that wearing a face mask can significantly reduce the transmission of respiratory viruses, including COVID-19. The masks work by filtering out respiratory droplets that may contain the virus, preventing them from reaching other people's noses and mouths. In conclusion, the image of the group of teenagers posing with their face masks highlights the importance of wearing PPE during the COVID-19 pandemic. The scientific evidence supporting the effectiveness of face masks in preventing the spread of respiratory viruses, including COVID-19, has been instrumental in promoting their use as a crucial public health measure.
Label	Positive
ALMT	Negative
Is hard sample?	✓
+ Context	Positive
+ Contextual Fusion	Positive
Sentence	Handsome young man standing in the kitchen with a cup of coffee at home.
Image	
Context	The image features a handsome young man standing in a kitchen with a cup of coffee. Recent studies have shown that moderate coffee consumption can have positive effects on cognitive function, memory, and mood. Additionally, coffee has been found to have potential health benefits, such as reducing the risk of certain diseases, including type 2 diabetes and certain types of cancer. In terms of advancements in coffee production and consumption, there has been a growing interest in sustainable and eco-friendly coffee practices. This includes the use of organic farming methods, fair trade practices, and the development of new technologies to reduce waste and improve the efficiency of coffee production.
Label	Neutral
LLaVA-v1.5	Neutral
Is hard sample?	✗
+ Context	Positive
+ Contextual Fusion	Neutral

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint* (2023). <https://arxiv.org/abs/2308.12966>
- [2] 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. <https://lmsys.org/blog/2023-03-30-vicuna/>
- [3] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *ICML*. <https://proceedings.mlr.press/v119/guo20h.html>
- [4] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *TMLR* (2022). <https://openreview.net/forum?id=jKN1pXi7b0>
- [5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS* 33 (2020). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. *arXiv preprint* (2023). <https://arXivpreprint.org/abs/2310.03744>
- [7] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *ICLR*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- [8] Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented Large Language Models with Parametric Knowledge Guiding. *arXiv preprint* (2023). <https://arxiv.org/abs/2305.04757>
- [9] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *NeurIPS* (2019). https://proceedings.neurips.cc/paper_files/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html
- [10] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint* (2023). arXiv:2303.08774 [cs.CL] <https://arXivpreprint.org/abs/2303.08774>
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.
- [12] K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *ICLR*. <https://arxiv.org/abs/1312.6034>
- [13] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *SIGIR* (, Virtual Event, Canada.) (*SIGIR '21*). Association for Computing Machinery, New York, NY, USA, 2443–2449. <https://doi.org/10.1145/3404835.3463257>
- [14] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In *EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). <https://doi.org/10.18653/v1/2023.emnlp-main.609>
- [15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS* 35 (2022).
- [16] Li Yang, Jianfei Yu, Chengzhi Zhang, and Jin-Cheon Na. 2021. Fine-grained sentiment analysis of political tweets with entity-aware multimodal network. In *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part I 16*. Springer. https://link.springer.com/chapter/10.1007/978-3-030-71292-1_31
- [17] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *EMNLP*. <https://aclanthology.org/2023.emnlp-main.49/>
- [18] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. *arXiv preprint* (2023). <https://arXivpreprint.org/abs/2309.07915>
- [19] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. In *Findings of ACL*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). <https://aclanthology.org/2023.findings-acl.519>