

# Supplementary Materials: Multimodal Intention Knowledge Distillation from Large Language Models for Social-Media Commonsense Discovery

Anonymous Authors

## 1 DISTILLATION PROMPT

After extracting image descriptions and key information from social media posts, we further refined and standardized the categorization of generated intentions, as depicted in Figure 1. We incorporated nine relations from ATOMIC [14], including xWant (After posting this tweet, the user wants to...), oEffect (Others will... after viewing this tweet), xAttr (The user posts this tweet because...), xIntent (The user intended to... by posting this tweet), xReact (The user feels... after posting this tweet), oReact (Others feel... after viewing this tweet), oWant (Others want to... after viewing this tweet), oEffect (After posting this tweet, the user will continue to...), and xNeed (Before posting this tweet, the user needs to...), along with an "Open" category for open-domain intentions. Here, "x" signifies the user's thoughts and behaviors post-posting, while "o" represents the impact on others. "Open" serves to elucidate the motivations and purposes behind a user's decision to publish specific content. By adopting this categorization method, we are able to comprehensively analyze posting intentions, accurately capture the underlying motives, and deepen our understanding of user behavior.

## 2 COMPARISON METHODS

### 2.1 Intention Distillation Baseline

To investigate whether different types of language models significantly impact intention generation without the MIKO framework, we empirically analyzed the plausible rate of generation across eleven large language models (LLMs): LLama2-7B[15], LLama2-13B[15], Mistral-7B-Instruct-v0.1[6], Mistral-7B-Instruct-v0.2[6], Falcon-7B[13], Flan-T5-xxl-11B[2], GLM3[4], GLM4[4], LLava-v1.5-13B[10], and LLava-v1.6-vicuna-7B[10].

**LLama2**[15] is an autoregressive language model using an optimized transformer architecture. The enhanced versions are fine-tuned through Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) to align more closely with human preferences for helpfulness and safety. Available configurations include 7B and 13B parameters.

**Mistral-7B-Instruct-v0.x**[6] comprises instructionally fine-tuned versions of the Mistral-7B model, using diverse public conversation datasets to improve the model's ability to understand and respond to instructions.

**Falcon-7B** is a 7B-parameter, causal, decoder-only model by TII, trained on 1,500 billion tokens from RefinedWeb, with additional curated corpora to enhance its generative capabilities.

**Flan-T5-xxl-11B**[2] is an improved version of the T5 model, fine-tuned on a mixture of tasks for significant performance enhancements, providing more sophisticated responses with the same parameter count.

**GLM**[4] describes a set of pre-trained language models based on an autoregressive fill-in-the-blank approach, with variants like GLM3 and GLM4 designed for specific language processing tasks.

**LLava**[10] is an open-source multimodal chatbot created by fine-tuning the LLama/Vicuna model with GPT-generated multimodal instruction-following data, leveraging an autoregressive transformer architecture for improved conversational capabilities.

### 2.2 Sarcastic Detection Baseline

In our study, we utilize both text-based and multimodal approaches as baseline frameworks to evaluate the impact of generated intentions. For text-based methods, we integrate **TextCNN** [7], **Bi-LSTM** [5], and **SMSD** [16]. Additionally, we adopt **BERT** [3], a robust baseline in sarcasm detection. In the multi-modal domain, our baselines encompass **HFM** [1], **D&R Net** [17], **Att-BERT** [12], **InCrossMGs** [8], **CMGCN** [9], and **HKE** [11].

**TextCNN**[7] applies a convolutional neural network (CNN) to text classification tasks, using multiple kernels of different sizes to extract key information from sentences (similar to n-grams of multiple window sizes), enabling it to better capture local correlations.

**Bi-LSTM** [5] utilizes two LSTM models, forward and backward, taking into account both past and future information, allowing the model to better capture contextual relationships in sequence data.

**SMSD** [16] employs self-matching networks and low-rank bilinear pooling for sarcasm detection.

**BERT** [3] serves as the competitive baseline for the sarcasm detection task.

**HFM** [1] treats text features, image features, and image attributes as three modalities and proposes a multimodal hierarchical fusion model to address the challenges of multimodal sarcasm detection.

**D&R Net** [17] is a novel method for modeling cross-modality contrast in the associated context, which models both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network (namely D&R Net). The decomposition network represents the commonality and discrepancy between image and text, and the relation network models the semantic association in a cross-modality context.

**Att-BERT** [12] is a BERT architecture-based model that focuses on both intra- and inter-modality incongruity for multimodal sarcasm detection. In detail, Att-BERT is inspired by the idea of a self-attention mechanism and design inter-modality attention to capturing inter-modality incongruity. Besides, the co-attention mechanism is applied to model the contradiction within the text. The incongruity information is then used for prediction.

**InCrossMGs** [8] utilizes an interactive graph convolution network (GCN) structure to jointly and interactively learn the incongruity relations of in-modal and cross-modal graphs for determining the significant clues in sarcasm detection.

**Intention prompt**

Based on the information below, guess the intention of why the user post this information. Generate different intentions if possible. The information are as follows:\\

Text: <post information>. \\

Image description: <image description>. \\

Concept: <concept information>. \\

Action: <action information>. \\

Object: <object information>. \\

Emotion: <emotion information>. \\

Keywords: <five related keyword>. \\

You can think about the **concepts, actions, object, emotions, and keywords**. Make the intention human-centric, and generate ten intentions by formulating your answer as:\\

Intention1: After posting this Tweet, the user wants to ...\\

Intention2: After viewing this Tweet, others will ...\\

Intention3: The user posts this Tweet because the user is ...\\

Intention4: The user posts this Tweet because the user intended to ...\\

Intention5: After posting this Tweet, the user feels ...\\

Intention6: After viewing this Tweet, others feel ...\\

Intention7: After viewing this Tweet, others want to ...\\

Intention8: After posting this Tweet, the user will ...\\

Intention9: Before posting this Tweet, the user needs to ...\\

Intention10: The user post this tweet because...

**Figure 1: An example illustrates the instructions for the intention generation.**

CMGCN[9] first detects objects paired with image modality descriptions, uses the description of the objects as a bridge, determines the importance of the association between the objects of the image modality and the context words of the text modality, and constructs a cross-modal graph. Subsequently, cross-modal graph convolutional networks are used to understand the inconsistent relationships between modalities for multi-modal sarcasm detection.

**HKE** [11] signifies a hierarchical framework, leveraging both atomic-level congruities through a multi-head cross-attention mechanism and composition-level congruity via graph neural networks, while a post exhibiting low congruity is identified as sarcastic.

### 3 MORE CASE STUDY

We show more examples of a raw text-image pair and their corresponding knowledge as well as image descriptions (Section 4.1), key information (Section 4.2), and different aspects of generated intentions (Section 4.3) in Figure 2. Similarly, plausibility and typicality are used to measure the quality of generated information, and we can observe that most generated intentions are both reasonable and comprehensive, in line with human intuitive understanding.


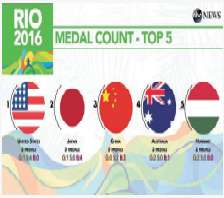
Social Media Post	Key-information	Intentions	P	T
<p>Just because you don't see it in the pool doesn't mean you're not sweating @GatoradePH #NothingBeatsGatorade</p> 	<p><b>Image description:</b> The image features a swimming pool with several people swimming in it. There are also a few sports bottles placed on the side of the pool, likely for the swimmers to stay hydrated. The bottles are lined up in a container, and they are of the Gatorade brand. The scene suggests that the swimmers are taking a break or preparing to swim, and the Gatorade bottles are there to provide them with the necessary fluids to keep them energized and refreshed during their swimming session.</p> <p><b>Concept:</b> Hydration. <b>Action:</b> Providing. <b>Object:</b> Gatorade bottles. <b>Emotion:</b> Refreshment. <b>Keywords:</b> Swimming, pool, sports, energized, hydrated.</p>	<p><i>xWant:</i> After posting this Tweet, the user wants to remind people about the importance of staying hydrated even when swimming. <i>oEffect:</i> After viewing this Tweet, others will be reminded to bring Gatorade for their next swimming session. <i>xAttr:</i> The user posts this Tweet because the user is promoting Gatorade as a refreshing drink for swimmers. <i>xiIntent:</i> The user posts this Tweet because the user intended to raise awareness about the need for proper hydration during physical activities. <i>xReact:</i> After posting this Tweet, the user feels satisfied that they have shared helpful information. <i>oReact:</i> After viewing this Tweet, others feel motivated to drink Gatorade for their next sports activity. <i>oWant:</i> After viewing this Tweet, others want to try Gatorade for their next swimming session. <i>oEffect:</i> After posting this Tweet, the user will continue to promote Gatorade as a drink for active individuals. <i>xNeed:</i> Before posting this Tweet, the user needs to ensure that the image and caption are visually appealing and informative. <i>Open:</i> The user post this tweet because they want to establish Gatorade as a go-to drink for people who engage in physical activities like swimming.</p>	H	H
<p>I just met @neiltyson. Had a total meltdown. Still in disbelief.</p> 	<p><b>Image description:</b> In the image, a man and a woman are standing next to each other, both wearing winter clothing. The man is wearing a cowboy hat, while the woman is wearing a scarf. They appear to be posing for a picture together. The man is smiling, and the woman is also smiling, indicating that they are enjoying the moment. The woman is wearing a coat, and the man is wearing a jacket. The scene suggests that they might have met someone famous, as the woman is described as having a meltdown and being in disbelief.</p> <p><b>Concept:</b> Meeting a famous person. <b>Action:</b> Posing for a picture. <b>Object:</b> Winter clothing. <b>Emotion:</b> Overwhelmed. <b>Keywords:</b> Meltdown, disbelief, smiling, famous, coat, jacket, cowboy hat, scarf.</p>	<p><i>xWant:</i> After posting this Tweet, the user wants to share their excitement and experience of meeting a famous person with their followers. <i>oEffect:</i> After viewing this Tweet, others will be curious about who the user met and share their own experiences of meeting famous people. <i>xAttr:</i> The user posts this Tweet because the user is amazed and overwhelmed by meeting someone they admire and want to share their experience with others. <i>xiIntent:</i> The user posts this Tweet because the user intended to express their gratitude for the opportunity to meet a famous person and document the moment. <i>xReact:</i> After posting this Tweet, the user feels relieved and happy to have shared their experience with others. <i>oReact:</i> After viewing this Tweet, others feel happy for the user and inspired to meet their own idols. <i>oWant:</i> After viewing this Tweet, others want to ask the user questions about their experience and who they met. <i>oEffect:</i> After posting this Tweet, the user will check their notifications and respond to any comments and messages from followers. <i>xNeed:</i> Before posting this Tweet, the user needs to select the best image and write a caption that accurately conveys their emotions. <i>Open:</i> The user posts this tweet because they want to show off their winter clothing and the celebrity they met.</p>	H	H
<p>Americans: fixing the medal tally in their favour since the #Olympics? were invented. IT'S GOLD FIRST #Rio2016?</p> 	<p><b>Image description:</b> The image features a graphic with a list of countries and their respective medal counts. The countries are represented by flags, with the United States, China, Australia, Japan, and Great Britain being the top five. The flags are arranged in a descending order, with the United States having the most medals, followed by China, Australia, Japan, and Great Britain. The graphic is likely meant to showcase the medal count at the Rio 2016 Olympics, with the United States being the top performer.</p> <p><b>Concept:</b> Olympic medal tally. <b>Action:</b> Fixing. <b>Object:</b> Medal count. <b>Emotion:</b> Pride. <b>Keywords:</b> Americans, Gold, Rio 2016, United States, China, Australia, Japan, Great Britain, Flags, Top performer.</p>	<p><i>xWant:</i> After posting this Tweet, the user wants to show pride in the United States' performance at the Rio 2016 Olympics. <i>oEffect:</i> After viewing this Tweet, others will feel impressed by the United States' dominance in the medal tally. <i>xAttr:</i> The user posts this Tweet because the user is a fan of the United States Olympic team and wants to celebrate their success. <i>xiIntent:</i> The user posts this Tweet because the user intended to start a conversation about the United States' dominance in the medal count. <i>xReact:</i> After posting this Tweet, the user feels patriotic and proud of the United States. <i>oReact:</i> After viewing this Tweet, others feel surprised by the United States' overwhelming success in the medal tally. <i>oWant:</i> After viewing this Tweet, others want to learn more about the individual events and athletes that contributed to the United States' medal count. <i>oEffect:</i> After posting this Tweet, the user will receive engagement from other users who want to discuss the Olympic results. <i>xNeed:</i> Before posting this Tweet, the user needs to research the medal count and order the flags in the graphic correctly. <i>Open:</i> The user post this tweet because they want to call out any accusations of the United States cheating or fixing the medal tally in their favor.</p>	L	L

Figure 2: The additional examples provided showcase the generated descriptions of images, key information, and their associated intentions. "P" denotes plausibility, while "T" represents typicality. Generated outcomes of high quality are accentuated in green, whereas those of lower quality are underscored in red. Furthermore, "H" signifies high, and "L" signifies low, referring to the respective scores for plausibility and typicality.

## REFERENCES

- [1] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *ACL*. 2506–2515.
- [2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, 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. <https://doi.org/10.48550/ARXIV.2210.11416>
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). 4171–4186.
- [4] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *ACL*. 320–335.
- [5] Alex Graves and Jürgen Schmidhuber. 2005. Frameworkwise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.
- [6] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023).
- [7] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. 1746–1751.
- [8] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs. In *MM*. 4707–4715.
- [9] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In *ACL*. 1767–1777.
- [10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. *CoRR* abs/2310.03744 (2023).
- [11] Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. In *EMNLP*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). 4995–5006.
- [12] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *Findings of EMNLP*, Trevor Cohn, Yulan He, and Yang Liu (Eds.), Vol. EMNLP 2020. 1383–1392.
- [13] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023). arXiv:2306.01116
- [14] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*. 3027–3035.
- [15] 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023).
- [16] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling. In *WWW*, Ling Liu, Ryan W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). 2115–2124.
- [17] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association. In *ACL*. 3777–3786.