WHY FOUNDATION MODELS STRUGGLE WITH CROSS-MODAL CONTEXT

Chen Henry Wu* **Neil Kale*** **Aditi Raghunathan**
Carnegie Mellon University
{chenwu2,nkale,aditirag}@cs.cmu.edu

ABSTRACT

Foundation models (FMs) deployed in real-world tasks – such as computer-use agents – must integrate diverse modalities. How good are FMs at performing joint reasoning over cross-modal context? To better understand this problem, we study FMs on *cross-modal conflicts*: scenarios where conflicting evidence is presented across modalities. This allows us to examine whether FMs prioritize one modality over another or reason jointly to reconcile the conflict. Our experiments reveal that FMs can recognize conflicts in unimodal contexts 90% of the time, but the ratio falls as low as 3% when evidence is split across modalities – similar observations hold in cross-lingual settings. We trace this failure to *cross-modal attention imbalance*, showing that FMs exhibit extreme asymmetry in attention scores, disproportionately prioritizing certain modalities. We show that cross-modal attention imbalance does not go away by simply scaling up multimodal or multilingual datasets blindly, since they lack training examples that explicitly require cross-modal reasoning. We demonstrate that even a simple and scalable method of explicitly combining multiple modalities within each training instance significantly reduces attention imbalance. Our findings underscore the importance of systematically addressing cross-modal contexts to build reliable FMs.

Figure 1: FM-based agents need to reason over diverse modalities, such as multilingual news, online shopping websites, maps, and EHR records. Failure to handle cross-modal context can result in consequences including misinformation (orange), purchasing a scam (yellow), misdirection (blue), or even providing the wrong medical treatment (light blue).

1 INTRODUCTION

Recent advances in foundation models (FMs; OpenAI, 2024a; Gemini, 2024; Anthropic, 2024b) have enabled their deployment in increasingly complex tasks that require reasoning over diverse information sources. From autonomous web browsing (Adept, 2022; Anthropic, 2024a; OpenAI, 2024b) to AI-driven research assistants (Perplexity, 2024; Sakana, 2024), FMs are now tasked with reasoning jointly over multiple domains such as text, images, code, and structured data.

*Equal contributions

However, existing work indicates that FMs fall short when handling inputs from non-textual modalities. For example, some studies show that FMs answer visual questions primarily based on language priors, disregarding visual inputs (Winterbottom et al., 2020a; Niu et al., 2020; Lin et al., 2024); others illustrate scenarios where models hallucinate objects absent from the image during open-ended generation tasks (Sun et al., 2023). Yet, it remains unclear whether this behavior originates primarily from a context-parametric gap (Goyal et al., 2024) or a modality gap (Liang et al., 2022).

In this work, we specifically focus on the capability of FMs to reason across modalities when all necessary information is explicitly provided in the input context. This setup is especially relevant for FM agents and assistants that need to interpret up-to-date information unavailable within their parametric knowledge – such as web pages combining images, multilingual text, and embedded scripts (Figure 1). By designing scenarios that isolate cross-modal reasoning from parametric knowledge retrieval, we directly assess how effectively these models reason over multiple modalities.

As a clean and concrete test case for this, we create *cross-modal conflict* datasets where each modality provides contrasting evidence. This allows us to examine whether FMs prioritize one modality over another or reason jointly to reconcile the conflict. Our experiments reveal a striking gap: while FMs perform well in *unimodal* settings (e.g., text-text or image-image), their ability to detect conflicts deteriorates significantly by up to 65% in *cross-modal* contexts (e.g., text-image). Moreover, this degradation extends to multilingual scenarios, where monolingual performance (e.g., English-English or Chinese-Chinese) is significantly better than multilingual performance (e.g., English-Chinese).

We investigate what drives this behavior, where state-of-the-art models exclusively rely on evidence from one modality rather than jointly reasoning. First, we observe that it is not simply a consequence of models being weak in one modality (§2). VLMs detect conflicts between multiple images as easily as conflicts between multiple texts (Figure 3). This extends to multilingual settings – conflicts between multiple Chinese texts are detected as often as in English.

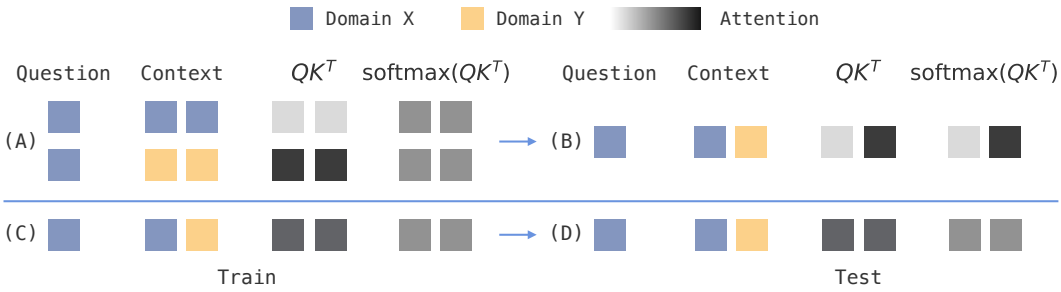

Figure 2: **An illustration for cross-modal attention imbalance.** In unimodal contexts (A), different domains show balanced normalized attention ($\mathrm{softmax}(QK^\top)$) despite divergent pre-softmax logits ($QK^\top$). Cross-modal contexts (B) expose cross-modal attention imbalance – normalization fails to mitigate logit-level imbalance. Instance-level modality mixing (C) resolves this by training models to intrinsically balance attention logits across modalities.

We hypothesize that the gap between unimodal and cross-modal conflict detection is because of *cross-modal attention imbalance*: an extreme asymmetry in attention scores, where FMs disproportionately prioritize certain modalities. We validate our hypothesis by finding that manual attention reweighting vastly shifts the model towards joint reasoning rather than relying on one modality over another (§3).

We investigate how to correct cross-modal attention imbalance. The problem is not resolved by simply adding more training data in each modality (§4.1). As illustrated in Figure 2, when the cross-modal attention scores are imbalanced, different modalities have different pre-softmax logits ($QK^\top$). However, after normalization, unimodal contexts show balanced normalized attention ($\mathrm{softmax}(QK^\top)$) and their performance remains stable. So, fine-tuning on either modality separately does not reduce attention imbalance.

Current instruction-tuning datasets do not involve joint reasoning over multiple modalities. This is a known problem — curating data for non-textual modalities is expensive (Liu et al., 2023; Dai et al., 2023). It is infeasible to curate large amounts of joint reasoning data on top of the instruction data from each modality. However, we hypothesize that correcting for cross-modal attention imbalance is already sufficient to promote joint reasoning. A simple and scalable way to do this is to simply

concatenate instructions from multiple modalities within the same context. In other words, we can repurpose existing datasets with this twist to greatly improve cross-modal joint reasoning in FMs.

In summary, we uncover a new fundamental gap between modalities – in terms of how they are processed in context. We demonstrate that state-of-the-art models fail in a simple cross-modal reasoning task of handling conflicting evidence from multiple modalities. We trace this failure to an imbalance in attention weights across modalities that can be addressed simply by mixing existing instruction data to create cross-modal instructions. Our findings also generally highlight the need for training paradigms that mirror the real-world complexity faced by FMs.

2 STRESS-TESTING CROSS-MODAL REASONING

We study the free-form generation from a FM. The FM takes a context C and a question Q as input and samples a response $y \sim \text{FM}(C, Q)$. The context C has an in-context knowledge conflict, i.e., C contains two subsequences, C_1 and C_2, that support contradictory answers to Q. We consider (C, Q) to be a unimodal conflict if $C_1, C_2 \in M_1$ and a cross-modal conflict if $C_1 \in M_1, C_2 \in M_2$ where M_1, M_2 are distinct modalities. This allows us to examine whether FMs prioritize one evidence over another or reason jointly to reconcile the conflict. Given a set of context-question pairs $\mathcal{D} = \{(C_i, Q_i)\}_{i=1}^{N}$, we define the *conflict detection rate* as the proportion of samples that are mentioned to contain conflicts. We used GPT-4o as the evaluator (see prompts in §E.1). To isolate the context-based reasoning independent of the parametric bias, we focus on tasks that depend on the context and cannot be solved with the parametric knowledge alone.

We construct two datasets: a *cross-modal question answering (CMQA)* dataset and a *cross-lingual question answering (CLQA)* dataset, each with controlled variations in context. The construction pipeline and examples from both data sets are given in §C.

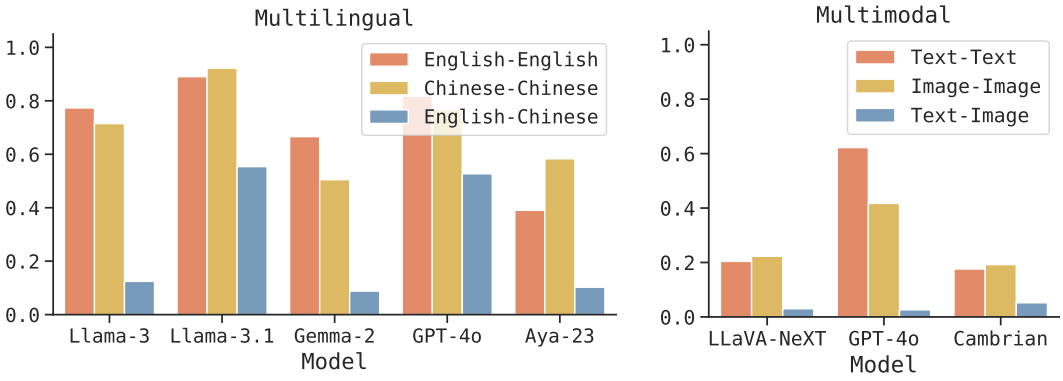

Figure 3: FMs are worse at reasoning over cross-modal contexts than unimodal contexts.

Figure 3 shows the performance of the FMs on our CLQA and CMQA datasets. We see that the conflict detection performance is significantly lower with the cross-modal contexts than with the unimodal contexts. For CLQA (Figure 3 left), we the performance on `English-English` is comparable to `Chinese-Chinese`, and both are far better – up to 5x – than `English-Chinese`. This shows that the lower performance in the multilingual setting is not due to the limited general capability of the FM in Chinese. Also, recall that the questions are always in English, including in the `Chinese-Chinese` setting, so the lower performance with multilingual contexts is neither due to the language barrier between English and Chinese. Results for Turkish and Icelandic are similar to those for Chinese, so we put them in §G for conciseness. We see similar trends on the CMQA task (Figure 3 bottom) – the performance with unimodal contexts (`Text-Text` and `Image-Image`) is far better than the performance with cross-modal contexts (`Text-Image`) for all FMs.

Summary. State-of-the-art FMs fail in a simple cross-modal reasoning task of handling conflicting evidences in multiple modalities.

3 CROSS-MODAL ATTENTION IMBALANCE IN FMS

To investigate the mechanisms underlying this failure, we probe the context contribution in FMs. Most state-of-the-art FMs are autoregressive – at each step, the FM predicts the next token based on the context so far. For architectures like Transformers (Vaswani et al., 2017), the representation at each step can be decomposed into a linear combination of the contributions of each span of context. For example, in a Transformer FM, the output of an attention head in a layer at step t is defined as: $\boldsymbol{a}_t = \boldsymbol{W}_O \sum_{j=1}^{t} w_{t,j} \boldsymbol{v}_j$, where $\boldsymbol{v}_j$ is value output of the j-th token in the context, $w_{t,j}$ is the attention weight from the t-th token to the j-th token, and $\boldsymbol{W}_O$ is the output projection matrix. We can group tokens in the context based on their domain: $\mathcal{C}_k$ contains all token indices of the k-th group. We can rewrite $\boldsymbol{a}_t$ as:

$$\boldsymbol{a}_t = \sum_{k=1}^{K} \left(\sum_{j \in \mathcal{C}_k} w_{t,j} \boldsymbol{W}_O \boldsymbol{v}_j \right) := \sum_{k=1}^{K} \boldsymbol{u}_k. \tag{1}$$

The term $\boldsymbol{u}_k$ is a vector that the k-th context writes to the residual stream at step t. It shows that the context representation is a linear combination of each context's contribution.

We hypothesize that the context contribution *in the task-relevant subspaces* is imbalanced in the cross-modal contexts, making the FM more likely to rely on the dominant context instead of doing conflict detection. In Figure 2, we illustrate our mental model of attention imbalance. In unimodal contexts (A), different domains show balanced normalized attention ($\mathrm{softmax}(QK^\top)$) despite divergent pre-softmax logits ($QK^\top$). Cross-modal contexts (B) expose cross-modal attention imbalance – normalization fails to mitigate logit-level imbalance. Instance-level modality mixing (C) resolves this by training models to intrinsically balance attention logits across co-occurring domains.

To demonstrate attention imbalance, we compute the average norm of $\boldsymbol{u}_k$ for each context, averaged over all layers and attention heads.[1] Figure 4 shows that, for cross-lingual, the English context contributes more than the Chinese context; for cross-modal, text contributes more than images.

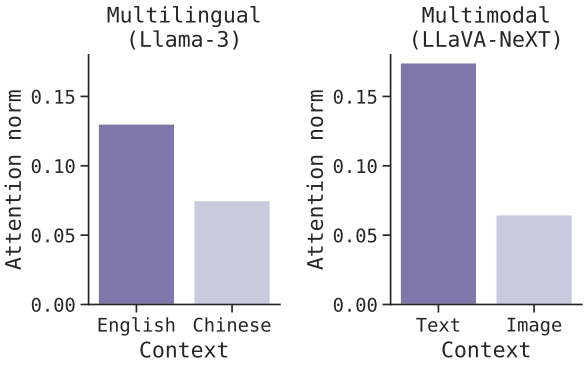

Figure 4: Cross-modal attention imbalance. English has larger attention than Chinese and images.

To test if there is a *causal* relationship between attention imbalance and cross-modal reasoning, we causally intervene the contribution of a context $\mathcal{C}_k$ by adding a small constant ϵ to its unnormalized attention score. Formally, denote the normalized attention weights at step t as $\boldsymbol{w}_t := [w_{t,1}, \ldots, w_{t,t}]^\top$. We manipulate the attention weights as follows:

$$\mathrm{Manip}(\boldsymbol{w}_t) = \mathrm{softmax}\left(\log \boldsymbol{w}_t + \epsilon \mathbf{1}_{\mathcal{C}_k}\right), \tag{2}$$

where $\mathbf{1}_{\mathcal{C}_k}$ is a vector with 1's on all the $\mathcal{C}_k$ context positions and 0's otherwise.

Figure 5 shows that the conflict detection performance indeed improves after attention manipulation. In the cross-lingual setting, the absolute improvement is up to 43% (relative by 5x). In the cross-modal setting, we observe a smaller yet significant gain of 18%. We hypothesize that this is because current visual instruction tuning datasets (Liu et al., 2024a) mainly focus on questions that can be answered (e.g., questions about an object in the image), which creates a strong bias towards responding with a definite answer. As a side observation, we find that attention manipulation can help the unimodal context as well: we find that FMs exhibit primacy bias and tend to rely more on the context that appears first. By increasing the attention weights on later context, we also further improve the conflict detection performance on unimodal context.

[1] We note that $\boldsymbol{u}_k$ averaged over all layers and attention heads should be viewed as a proxy of what we want to measure, i.e., context contribution *in the task-relevant subspaces*. We further discuss this in §F. For this reason, we do not argue that the norms of different $\boldsymbol{u}_k$ should be the same to achieve the best performance.

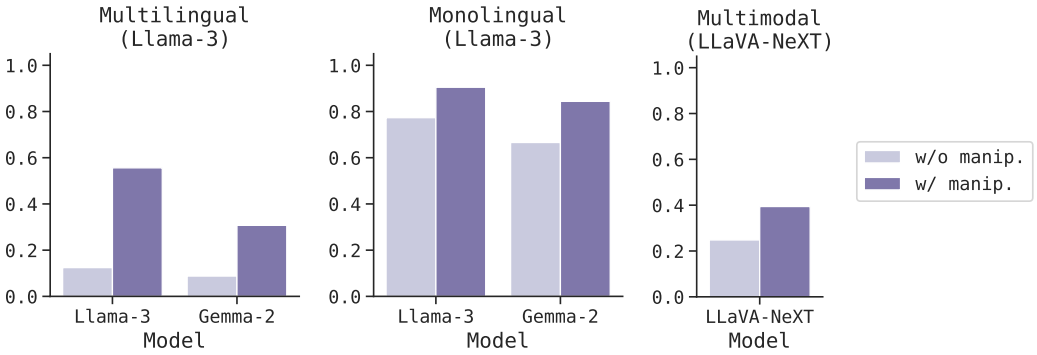

Figure 5: Cross-modal attention imbalance has a causal effect on cross-modal reasoning: we apply a fixed attention bias to increase the attention over the context with a smaller attention output norm, and see that this improves the conflict detection performance. We use the `Standard` prompt for cross-lingual and monolingual settings, and `Explicit` prompt for the cross-modal setting.

Summary. Cross-modal attention imbalance has a causal negative effect on FMs' cross-modal reasoning capability.

4 How to correct cross-modal attention imbalance?

4.1 Dataset-level modality mixing does not help

We begin by noting that most state-of-the-art FMs today are trained on highly diverse corpora, spanning a wide range of domains and multiple languages (Meta, 2024; Riviere et al., 2024). More surprisingly, as we observe in Figure 3, Aya-23, a FM specifically optimized for multilingual capabilities, performs no better than other FMs with multilingual contexts. This suggests that simply training FMs on diverse modalities does not, by itself, ensure good cross-modal reasoning.

To reinforce this, we run two instruction-tuning experiments. First, we finetune Llama-3 on mixed English and Chinese instruction tuning datasets (we call this strategy **dataset-level modality mixing**) and see if this improves conflict detection in the cross-lingual `English-Chinese` setting. Specifically, we use the English and Chinese subsets of Bactrian-X (Li et al., 2023), a multilingual instruction-tuning dataset containing 67k samples in each language. Similarly, we finetune Qwen-2.5-VL on mixed text and visual instruction tuning datasets (**dataset-level modality mixing**) and see if this improves conflict detection in the cross-modal `text-image` setting. In this experiment, we use the visual instruction data from Liu et al. (2024a) and the English subset of Bactrian-X. In Figure 6, we see that *dataset-level modality mixing* offers minimal gains in alleviating cross-domain attention imbalance. This motivates us to understand why diverse, multimodal data is not enough to close the gap between unimodal and cross-modal contexts.

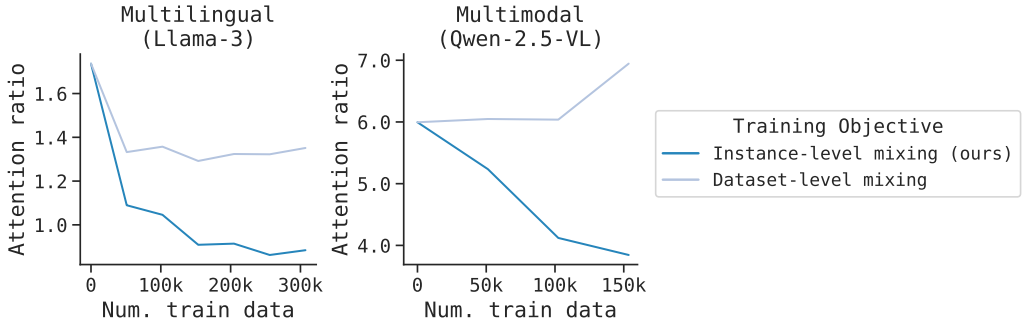

Figure 6: Finetuning on instance-level mixed data (dark blue) reduces cross-modal attention imbalance significantly more than fine-tuning on dataset-level mixed data (gray).

4.2 INSTANCE-LEVEL MODALITY MIXING

We have shown that standard cross-modal instruction tuning (e.g. having both English and Chinese examples in the data, or having both text and visual instruction tuning examples) fails to improve cross-modal attention imbalance. We hypothesize that the gap in unimodal and cross-modal contexts arises because mixing datasets does not expose models to instances requiring cross-domain reasoning *within the same context*. Without instance-level modality mixing between modalities, the pre-softmax attention scores for one domain could be hugely different from that of another domain, without changing the normalized attention scores on each domain (Figure 2). To address the lack of instance-level modality mixing between modalities, we propose a simple and scalable method of explicitly combining modalities within each training instance. For example, in the cross-lingual setting:

Input:
`<Chinese instruction> <English instruction>`
`Reply to both user instructions.`
Output:
`<Chinese response> <English response>`

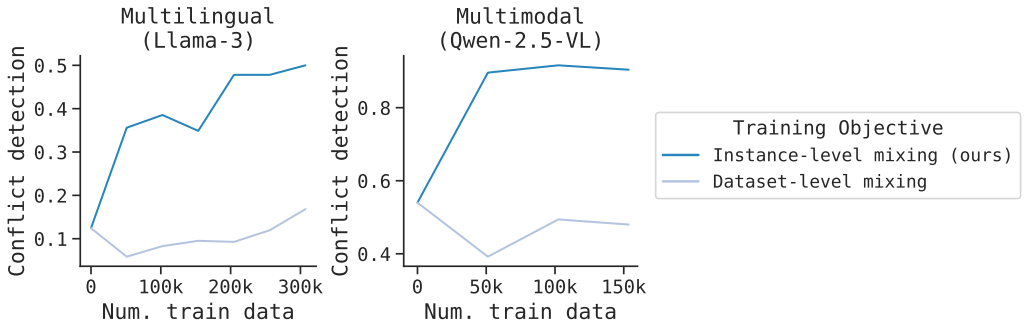

Figure 7: Finetuning on instance-level mixed data (dark blue) improves cross-modal conflict detection largely more than fine-tuning on traditional dataset-level mixed data (gray).

To verify the benefit of instance-level modality mixing, we use the same data from § 4.1 but mix them at the instance level instead of at the dataset level. In Figure 6, we report the attention imbalance of model checkpoints for instance-level modality mixing and the dataset-level modality mixing baseline in §4.1. In the cross-lingual setting, instance-level modality mixing reduces attention imbalance between modalities by $4\times$. In the cross-modal setting, it reduces attention imbalance by 34%. In Figure 7, we report the performance of model checkpoints for *instance-level modality mixing* and the baseline in §4.1 (*dataset-level modality mixing*). In the cross-lingual setting, instance-level modality mixing boosts conflict detection by 37%, much greater than dataset-level modality mixing. In the cross-modal setting, instance-level modality mixing improves conflict detection by $2\times$.

Summary. Instance-level modality mixing mitigates attention imbalance and improves cross-modal reasoning, without requiring any additional data curation.

5 CONCLUSIONS

We uncovered a new fundamental gap in how FMs process modalities in context. Through controlled datasets and experiments, we demonstrated that FMs fail in a simple cross-modal reasoning task of handling conflicting evidences in multiple modalities. Our analyses trace the problem to *cross-modal attention imbalance*, an imbalance in attention weights across modalities. We showed that simply including multiple modalities in training (i.e., dataset-level modality mixing) has little gains, while explicitly mixing different modalities within each training sample (i.e., instance-level modality mixing) mitigates attention imbalance and substantially boosts conflict detection. Our results highlight the need for training paradigms that mirror the real-world complexity faced by models and for methods that enable foundation models to balance cross-modal attention and reason on cross-modal contexts.

REFERENCES

Adept. ACT-1: Transformer for actions, 2022.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. *Anthropic Blog*, 2024a.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. *Anthropic Blog*, 2024b.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. In *OpenAI*, 2023.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. *ArXiv*, abs/2205.10964, 2022.

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

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Gemini. Introducing gemini 2.0. *Blog*, 2024.

Sachin Goyal, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. *arXiv preprint arXiv:2410.10796*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 2016.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. *arXiv preprint arXiv:2105.01993*, 2021.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. *Blog*, 2024a.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*, 2022.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. In *ArXiv*, 2023.

Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, Yupeng Hou, Fuxiao Liu, and Tianyi Zhou. Mosaic-it: Free compositional data augmentation improves instruction tuning. *arXiv preprint arXiv:2405.13326*, 2024b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *ArXiv*, abs/2203.02053, 2022.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. Multi-modal contrastive representation learning for entity alignment. *arXiv preprint arXiv:2209.00891*, 2022.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *International Conference on Machine Learning*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024a.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Llama Team AI Meta. The llama 3 herd of models, 2024.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, and Jugal Kalita. The less the merrier? investigating language representation in multilingual models. *ArXiv*, 2023.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xiansheng Hua, and Ji rong Wen. Counterfactual vqa: A cause-effect look at language bias. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020.

OpenAI. Hello gpt-4o. *OpenAI Blog*, 2024a.

OpenAI. Introducing operator research preview. *OpenAI Blog*, 2024b.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023.

Perplexity. Pro search: Upgraded for more advanced problem-solving. *Perplexity Blog*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron,

Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, L. Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, S. Mc Carthy, Sarah Perrin, S'ebastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Brian Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Cl'ement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *ArXiv*, 2024.

Sakana. Ai scientists: Entering an era in which ai conducts its own research. *Sakana Blog*, 2024.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *arXiv preprint arXiv:2406.18406*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *ArXiv*, abs/2309.14525, 2023.

A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In *ACL*, 2024.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and N. A. Moubayed. On modality bias in the tvqa dataset. *ArXiv*, abs/2012.10210, 2020a.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210*, 2020b.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

X Yue, Y Ni, K Zhang, T Zheng, R Liu, G Zhang, S Stevens, D Jiang, W Ren, Y Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arxiv, 2023.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*, 2023.

A Related Work

Modality Gap in Foundation Models FMs are known to fall short when handling inputs from low-resource modalities. For example, some studies demonstrate that FMs answer visual questions primarily based on language priors, disregarding actual visual inputs (Winterbottom et al., 2020a; Niu et al., 2020; Lin et al., 2024); others illustrate scenarios where models hallucinate objects absent from the image during open-ended generation tasks (Sun et al., 2023). Yet, it remains unclear whether this behavior originates primarily from a context-parametric gap (Goyal et al., 2024) or a modality gap (Liang et al., 2022). Modern vision-language models (VLMs) embed text and images into a shared embedding space (Radford et al., 2021; Jia et al., 2021). The modality gap is characterized as separation between the embeddings of different data modalities (Liang et al., 2022) which hurts performance on visual question answering and classification tasks (Guo et al., 2023; Winterbottom et al., 2020b). Several explanations have been proposed, including inductive bias of encoders and disuniformity of contrastive loss (Fahim et al., 2024). A similar phenomenon persists in multilingual FMs too (Nigatu et al., 2023; Chang et al., 2022).

Multilingual and Multimodal Instruction Tuning Multilingual and vision-language models employ specialized pre-training data (Ustun et al., 2024; Li et al., 2024a) and instruction-tuning datasets (Li et al., 2023; Liu et al., 2023; Antol et al., 2015) to improve performance on underrepresented modalities. In general, however, these models are designed for unimodal performance and they do not saturate large cross-modal benchmarks like MMMU and ScienceQA that require simultaneously reasoning over data in multiple modalities (Yue et al., 2023; Lu et al., 2022). Several fine-tuning approaches have been suggested to improve cross-modal reasoning in FMs. For example, X-InstructBLIP claims that training on individual modalities can result in emergent cross-modal reasoning (Panagopoulou et al., 2023). In our analysis, however, we find strong evidence that this is not always possible.

Related Approaches for Cross-Modal Reasoning Another common approach to improve cross-modal performance is to mitigate language bias, or over-dependence on language priors (Niu et al., 2020). This approach prevents the model from ignoring images due to parametric knowledge about the question, but does not counteract bias within the context towards text over image evidence. Other more specialized approaches include aligning individual entities between modalities (Lin et al., 2022) or learning sparse feature representations that rely less on language priors (Guo et al., 2021). In addition, (Li et al., 2024b) propose a fine-tuning approach that is similar to our instance-level mixing strategy; however, they focus only on textual data, whereas we focus entirely on mixing modalities.

Knowledge Conflicts in Cross-Modal Contexts Even in unimodal settings, FMs sometimes fail to identify when they encounter conflicting information (Xu et al., 2024). In these unimodal scenarios (e.g., correcting outdated facts), there is evidence that FMs exhibit self-consistency – the ability to identify when they don't know an answer (Kadavath et al., 2022; Yin et al., 2023). In addition, mitigation strategies like prompting (Zhou et al., 2023), pretraining (Li et al., 2022), and reweighting neurons (Shi et al., 2024) are known to improve detection but remain limited to specific unimodal contexts. Existing instruction-tuning solutions for conflict detection (Wang et al., 2023) rely heavily on curated conflict-specific datasets. Notably, prior work largely overlooks knowledge conflicts between multiple modalities. (Liu et al., 2024b) benchmarks cross-modal conflicts, but focuses on context-parametric conflicts between images and the model's pretrained knowledge. To the best of our knowledge, knowledge conflicts have not previously been used to study cross-modal reasoning in a controlled setting.

B Additional examples of conflict detection

We provide additional examples of FM failure in conflict detection over cross-modal contexts.

Multilingual example In Figure 8, we provide an example of how an FM (GPT-4o) with web access can fail to acknowledge knowledge conflicts from multilingual news sources.

Cross-modal agent example In Figure 9, we provide an example of how an FM (GPT-4o) with web access can fail to acknowledge knowledge conflicts in multimodal product descriptions.

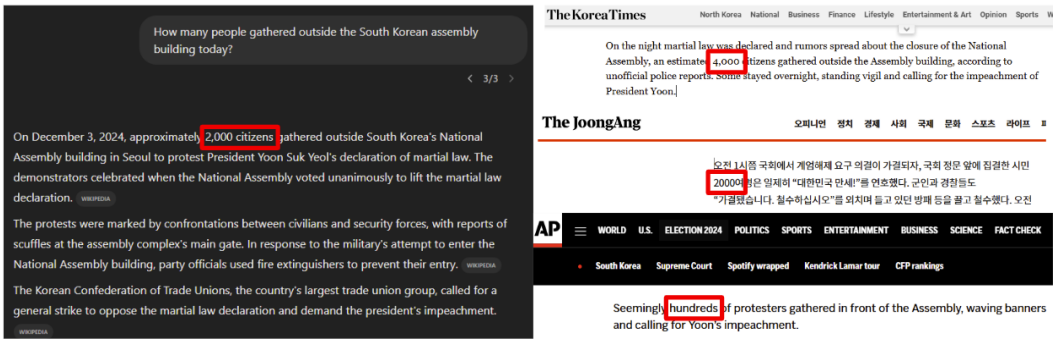

Figure 8: An FM with web access can fail to acknowledge knowledge conflicts from multilingual news sources. For example, GPT-4o reports the size of the protest outside South Korea's National Assembly on December 3, 2024 as 2,000 people, although different sources provide conflicting numbers of attendees.

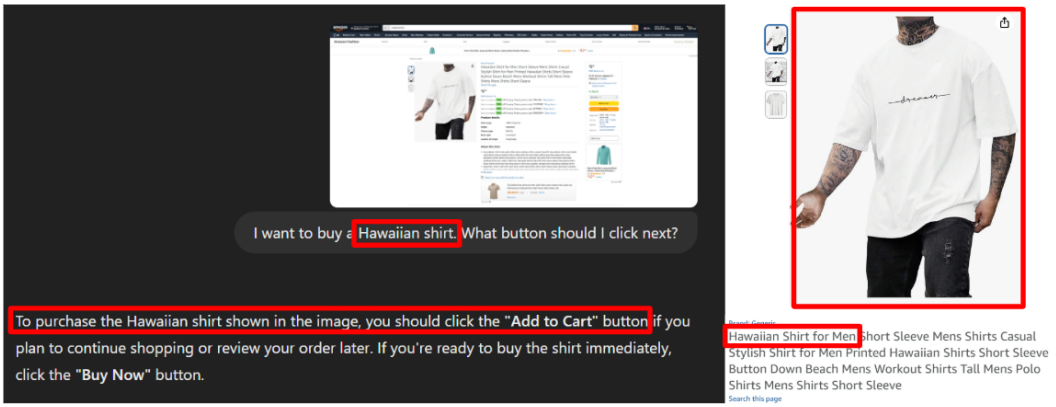

Figure 9: A FM can fail to acknowledge knowledge conflicts in multiple modalities. For example, GPT-4o instructs the user to purchase an item labeled as "Hawaiian Shirt for Men" despite the image clearly depicting an ordinary t-shirt, not a Hawaiian shirt.

C Dataset Details

C.1 Dataset Construction

Cross-modal question answering (CMQA) The multimodal question answering dataset is constructed over both image and text based on the VQA-v2 dataset (Goyal et al., 2016). Each sample in VQA-v2 consists of an image V, a question Q, and 10 candidate answers. In total, we subsample 500 triples of image, question, and answer (V, Q, A) from the dataset.

For each triplet, we prompt GPT-4o to generate a text description $\overline{T}$ that does not agree with the image V regarding the question Q, and the answer $\overline{A}$ based on $\overline{T}$. Given the image V, the text description $\overline{T}$, and the question Q, the FM should report a conflict as A is contradictory to $\overline{A}$. We name this dataset $\{(V, \overline{T}, Q, A, \overline{A})\}$ as `Text-Image`. For each image V, we prompt GPT-4o to generate a description T' that agrees with the image regarding the question Q. We name $\{(T', \overline{T}, Q, A, \overline{A})\}$ as `Text-Text`. For each $\overline{T}$, we prompt DALL-E 3 (Betker et al., 2023) to generate an image $\overline{V}$. We name $\{(V, \overline{V}, Q, A, \overline{A})\}$ as `Image-Image`.

Cross-lingual question answering (CLQA) We create a dataset of question answering over synthetic news paragraphs about fictitious events (so the FM cannot use parametric knowledge to answer the questions). We use GPT-4o to generate 400 topics. For each topic, we prompt GPT-4o to generate: (1) a synthetic news paragraph P_E in English which has not appeared in reality, a question Q in English, and an answer A based on the paragraph, and (2) synthetic news paragraph $\overline{P}_C$ in

Chinese that does not agree with the English one P_E regarding the question Q, and an answer $\overline{A}$ based on the Chinese one.

Given the two news paragraphs P_E and $\overline{P}_C$ and the question Q, the FM should reason over both since A is contradictory to $\overline{A}$. We name this cross-modal dataset $\{(P_E, \overline{P}_C, Q, A, \overline{A})\}$ as `English-Chinese`. We then derive several monolingual variants of different language combinations via (back-) translation. For each paragraph $\overline{P}_C$ in Chinese, we back-translate it into English $\overline{P}'_E$. We name $\{(P_E, \overline{P}'_E, Q, A, \overline{A})\}$ as `English-English`. For each English paragraph P_E, we translate it into Chinese P'_C. We name $\{(P'_C, \overline{P}_C, Q, A, \overline{A})\}$ as `Chinese-Chinese`. Similarly, we test other variants where Chinese is replaced with low-resource languages such as Turkish or Icelandic.

C.2 Dataset Examples

Figure 10 shows an example English evidence, Chinese evidence, and question from our CLQA dataset. Figure 11 shows an example text evidence, image evidence, and question from our CMQA dataset.

In a recent breakthrough, researchers at the University of California have developed a new genetic modification technique that significantly boosts crop yield. This advancement, based on CRISPR-Cas9 technology, allows for more precise editing of plant genomes, enabling scientists to enhance growth rates and resistance to environmental stressors. Preliminary field tests conducted on corn and wheat in the San Joaquin Valley have shown promising results. The modified crops exhibited a robust increase in productivity, with yield improvements recorded at approximately 25\%. This development is expected to revolutionize agricultural practices by reducing the need for chemical fertilizers and pesticides, potentially lowering production costs and benefiting farmers globally. According to Dr. Emily Zhang, the lead scientist on the project, the technique is ready for wider application and could be instrumental in addressing food security challenges posed by a growing global population and climate change. The next steps involve scaling up production and collaborating with agricultural organizations to implement these genetically modified crops on a larger scale. However, some environmental groups have raised concerns about the long-term impacts on biodiversity and ecosystem balance, advocating for more rigorous testing before widespread adoption.

最近，加州大学的研究人员开发了一种新的基因改造技术，可以显著提高农作物产量。该技术基于CRISPR-Cas9技术，允许更精确地编辑植物基因组，从而增强生长速度和对环境压力的抵抗力。在圣华金谷进行的玉米和小麦初步田间试验显示出令人鼓舞的结果。经过改造的作物表现出生产力的显著提高，产量提高约为15%。这一发展预计将通过减少对化肥和农药的需求来革新农业实践，可能降低生产成本并使全球农民受益。项目负责人张艾米博士表示，该技术已准备好进行更广泛的应用，并可能在应对由全球人口增长和气候变化带来的粮食安全挑战中发挥关键作用。接下来的步骤包括扩大生产规模，并与农业组织合作，在更大范围内实施这些转基因作物。然而，一些环保团体对生物多样性和生态系统平衡的长期影响表示担忧，呼吁在广泛采用之前进行更严格的测试。

Q: What is the estimated percentage increase in crop yield due to the new genetic modification technique?

Figure 10: English evidence, Chinese evidence, and question from our CLQA dataset.

D Additional experiments with prompt variants

In addition to the prompts above (denoted as `Standard` in Figure 12), which do not assume intervention from the user, we also test other prompts that encourage FM to detect the conflict. Specifically, we explore two types of prompts: (1) add an instruction that tells the FM to report the conflict if it finds any (denoted as `Instructed` in Figure 12); (2) embed the question into a yes-no question: "Would the answers to the question '$\{Q\}$' be the same based on the paragraphs in the context?" (denoted as `Explicit` in Figure 12). In Figure 12, we see that, although the overall conflict detection performance improves, the trend is similar to Figure 3 – the conflict detection performance is lower in the cross-modal contexts than in the unimodal contexts. In the next section, we explore why this is the case and try to improve this practically.

In the breathtaking expanse of a winter wonderland, the person is immersed in the art of snowboarding. As they glide effortlessly down the pristine slopes, their board carves precise arcs into the powdery snow, leaving behind a trail of skillful mastery. Clad in vibrant winter gear, they exhibit the perfect blend of agility and grace that is the hallmark of a seasoned snowboarder. Each twist and turn is a testament to their years of practice and passion for the sport. The snowy landscape stretches out endlessly, a canvas for the snowboarder's dynamic movements. With each jump and trick, they defy gravity, soaring briefly before landing with practiced ease. The sun shines brightly, reflecting off the snow and illuminating the snowboarder's path as they navigate the mountain with confidence. Every moment on the board is a dance with the elements, a thrilling experience that captivates both the participant and any fortunate observers. In this wintry realm, the snowboarder finds freedom and exhilaration in equal measure, making the most of every descent.

Q: What is the person doing?

Figure 11: Text evidence, image evidence, and question from our CMQA dataset.

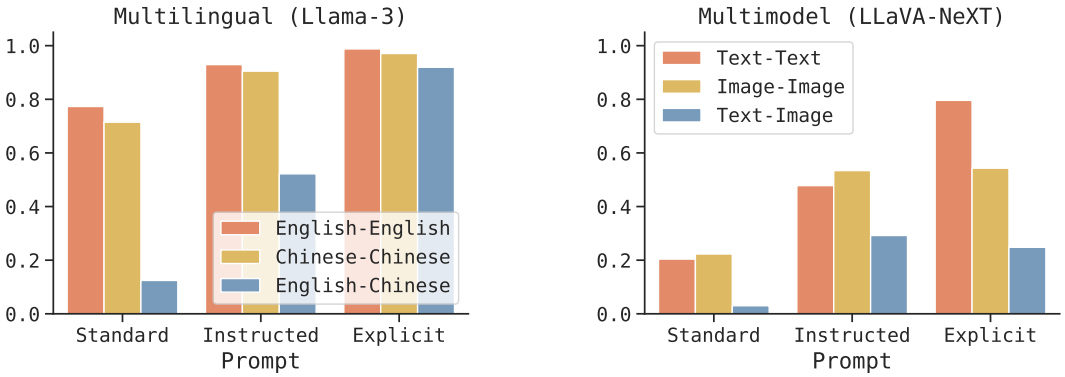

Figure 12: Ablation studies on the prompt. FMs are worse at reasoning over cross-modal contexts than unimodal contexts. See the text for details of each prompt.

E PROMPTS

E.1 LM-AS-A-JUDGE FOR EVALUATING CONFLICT DETECTION

We use the following prompt for GPT-4o to evaluate if the model output is doing conflict detection.

```
I'll provide you with a question and a response by a model.

## Your task

Can you infer from the response that the pieces of evidence provided to the model
has contradictions within them (or that different pieces of evidence suggest
different answers)?
If yes, then contradict_score = 1, otherwise contradict_score = 0.

Examples of contradictions:
- two images describing different things, e.g., one image describes something,
while the other image describes something else;
- image and text describing different things, e.g., the image describes something,
while the text describes something else;
- two paragraphs giving different answers, e.g., one paragraph says something,
while the other paragraph says something else.

## Input

Question: {question}
Response: {response}

## Output requirements

Wrap your final judgement in triple backticks. Your judgement should look like
this:

```json
{{
    "contradict_score": 0 or 1
}}
```
```

E.2 CMQA AND CLQA PROMPTS

We prompt FMs with two pieces of context and a question, and sample a response from the FM. Here is the prompt we use for our CLQA task:

```
Here are some paragraphs.
<paragraph_1>
<paragraph_2>
Based on all the paragraphs, answer the question below. Reply in English.
<question>
```

Similarly, we use the following prompt for our CMQA task:

```
<image> <text>   // The order depends on the model.
Above are visual and textual descriptions of a scene.
Answer the question below.
<question>
```

F ADDITIONAL EXPERIMENTS IN ATTENTION IMBALANCE

Recall that in §3 we demonstrate cross-modal attention imbalance with the average norm of $\boldsymbol{u}_k$ for each context, averaged over all layers and attention heads. In this section, we elaborate on this by visualizing $\boldsymbol{u}_k$ for each context in each layer and attention head.

Figure 13 visualizes the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. Figure 14 visualizes the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. We see that that the values over the Chinese/image context is generally smaller than those over the English/text context, especially in upper layers.

We note one important exception that is relevant to the footnote in §3, where we argue that $\boldsymbol{u}_k$ averaged over all layers and attention heads should be viewed as a proxy of what we want to measure, i.e., context contribution *in the task-relevant subspaces*. Figure 13, Layers 11-14 are an exception, where the values over the Chinese context is higher – we argue that these layers are *not* in the task-relevant subspace (i.e., they activates on the Chinese context but does not improve the reliance on Chinese when answering the question). For this reason, we do not argue that the norms of different $\boldsymbol{u}_k$ should be the same to achieve the best conflict detection performance.

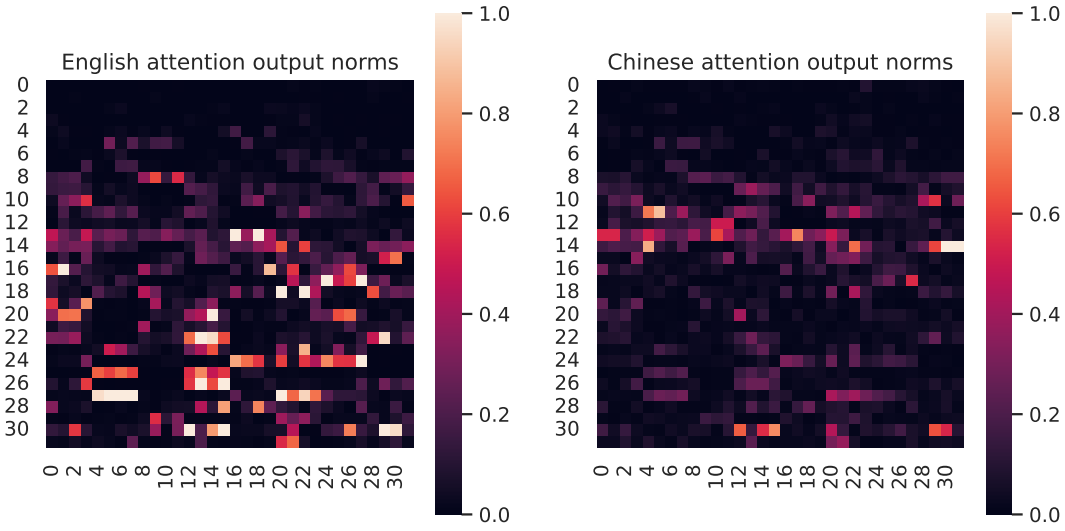

Figure 13: We visualized the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. We see that the values over the Chinese context is generally smaller than those over the English context, especially in upper layers. Notably, Layers 11-14 are an exception, where the values over the Chinese context is higher – we argue that these layers are *not* in the task-relevant subspace (i.e., they activates on the Chinese context but does not improve the reliance on Chinese when answering the question).

G ADDITIONAL RESULTS ON OTHER LANGUAGES

Figure 15 shows the results of conflict detection over cross-modal contexts containing Icelandic and Turkish.

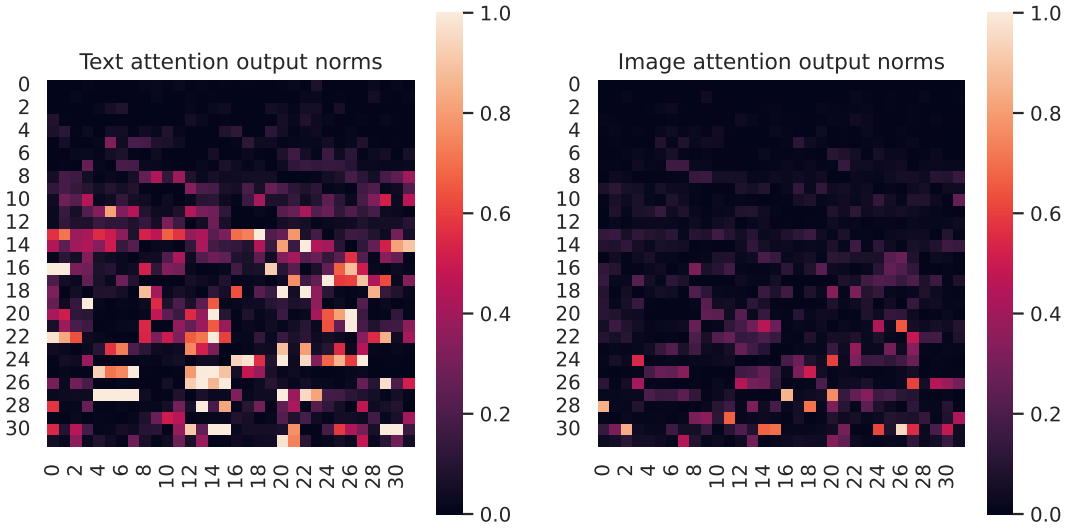

Figure 14: We visualized the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multimodal setting, aggregated over all test samples. We see that the values over the image is generally smaller than those over the text.

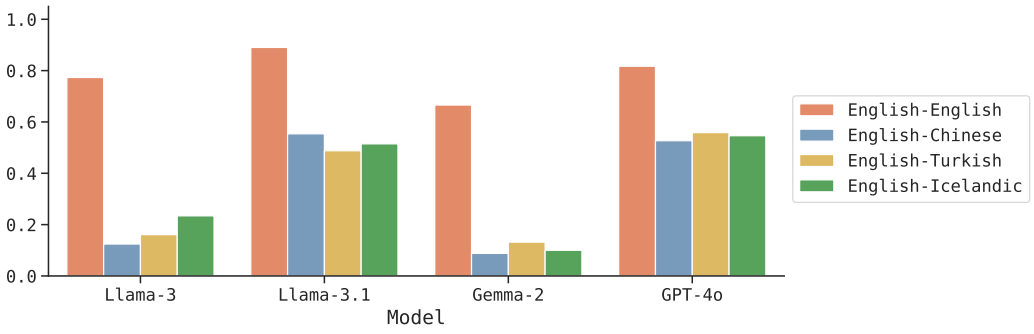

Figure 15: Conflict detection ratio over cross-modal contexts with Icelandic and Turkish.