CONTEXT HETEROGENEITY MAKES IN-CONTEXT KNOWLEDGE CONFLICTS HARDER TO DETECT

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models (LMs) deployed in real-world tasks – such as medical report synthesis, web navigation, and summarization – must process diverse inputs and handle conflicting information. Users expect them to detect *in-context knowledge conflicts* – direct contradictions about objective facts – and issue alerts. Yet, we find a critical failure: when faced with conflicting evidence across **heterogeneous contexts**, such as multiple languages or modalities, LMs fail to detect conflicts, leaving them vulnerable to attacks and misinformation. While they achieve near-perfect accuracy in homogeneous contexts, this **drops by up to 65%** in heterogeneous settings. We identify *context imbalance* as the root cause: LMs exhibit extreme attention asymmetry across domains, disproportionately prioritizing certain domains in mixed inputs. Current instruction-tuning, which trains on separate examples from multiple domains, fails to correct this. To address this, we need *instance-level diverse points* that require reasoning over multiple domains within a single context. We introduce **Heterogeneous Instruction-Tuning** (HeteroIT), a scalable dataset-mixing procedure that generates instance-level diversity by combining datasets from different domains. Applying our method to Bactrian-X, a standard multilingual instruction-tuning dataset, improves conflict detection by 37%.

1 INTRODUCTION

Figure 1: LM agents need to recognize inconsistency across domains, such as in multilingual news, shopping websites, maps, and EHR records. Failure to recognize conflicts can result in consequences ranging from purchasing a scam (yellow) to providing the wrong medical treatment (light blue).

Recent advances in language models (LMs; 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), LMs are now tasked with integrating text, images, code, and structured data from multiple sources. However, this expanded capability introduces new risks: models must determine what is true, safe, and relevant while navigating a landscape rife with misinformation, adversarial manipulation, and privacy threats.

For instance, a computer-use agent might summarize multilingual news reports on an ongoing election, process medical advice from conflicting sources, or purchase products from multimodal e-commerce sites (Figure 1). Yet, without robust mechanisms for conflict detection, these agents risk amplifying falsehoods, leaking private data, or executing malicious instructions. Worse, when interacting with external tools – such as APIs for financial transactions or system commands – models may blindly execute unsafe operations if they fail to recognize contradictions across diverse contexts.

To operate safely and reliably, LMs must process and reconcile *heterogeneous contexts* – inputs that mix different modalities and languages – such as web pages combining images, multilingual text, and embedded scripts. For example, in Figure 1(a), the LM must parse multilingual articles to answer a question about international events. Although the sources present conflicting information, the LM tends to rely on one of them. Similarly, in Figure 1(b), the LM must evaluate whether a scam product with inconsistent description and image is reliable to purchase, but the LM blindly trusts the deceptive description. The ability to identify and reason through conflicting evidence is critical, yet current models struggle to do so when information is distributed across multiple domains.

We systematically investigate conflict detection across controlled settings involving multiple languages and modalities, such as image and text. Our experiments reveal a striking gap: while LMs perform well in *homogeneous* settings, their ability to detect conflicts deteriorates significantly by up to 65% in *heterogeneous* contexts. This degradation occurs consistently across both multilingual and multimodal scenarios. Moreover, when conflicting information is present, models consistently favor one domain over the other.

Why do models that otherwise excel in individual domains, struggle so severely with conflict detection in heterogeneous settings? After all, state-of-the-art LMs are already trained on diverse datasets with trillions of multilingual tokens (Meta, 2024) and billions of images (Li et al., 2024). To further rule out lack of data diversity as the cause, we fine-tune Llama-3 on the Chinese and English subsets of Bactrian-X (Li et al., 2023a), a multilingual instruction-tuning dataset with 67k samples per language. Yet, this *dataset-level mixing* improves conflict detection on mixed English/Chinese evidence by only 4% – a negligible gain (Figure 6 (bottom)).

We investigate the model's attention mechanisms and identify *context imbalance* in heterogeneous contexts as the root cause of poor conflict detection. Despite being trained on multiple domains, models internally weigh them unevenly. For instance, when Llama-3 (Meta, 2024) answers questions based on mixed English/Chinese evidence, it ignores the Chinese context 87.5% of the time. Its attention layers exhibit severe domain bias – attention norms for English tokens dwarf those for Chinese by 2–3× (Figure 4), effectively muting cross-domain contradictions. We verify that training on diverse languages, as before, does little to reduce this imbalance (Figure 6 (top)). However, directly rebalancing attention across domains – without any training – boosts the performance by 43%.

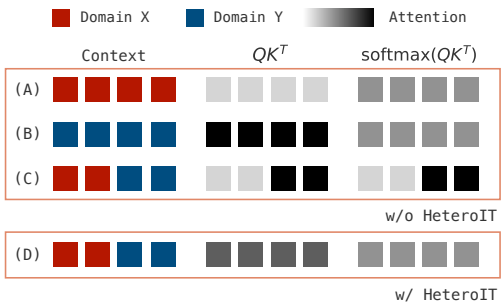

Figure 2: **Illustration for context imbalance.** In homogeneous contexts, i.e., (A) and (B), different domains show balanced normalized attention ($\mathrm{softmax}(QK^\top)$) despite divergent pre-softmax logits ($QK^\top$). Heterogeneous contexts (C) expose domain bias – normalization fails to mitigate logit-level imbalance. HeteroIT (D) resolves this by training models to intrinsically balance attention logits across domains.

When domain tagging at the input level is feasible and white-box access to the model is available, direct attention manipulation can help mitigate imbalance. However, a more scalable and practical solution is needed. We hypothesize that the gap between homogeneous and heterogeneous settings arises because simply mixing domains does not expose models to instances requiring cross-domain reasoning *within the same context.* In the absence of such instance-level diverse examples, attention scores for one domain could be hugely different from that of another domain, without hampering the performance on either domain (Figure 2).

To address this, we propose **Heterogeneous Instruction-Tuning** (HeteroIT), a new procedure that takes any existing datasets from different domains (e.g., English/Chinese instruction-response pairs from Bactrian-X) and combines them to create instance-level diverse examples, such as:

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

Fine-tuning Llama-3 on 67k HeteroIT-processed examples from Bactrian-X for just one epoch reduces context imbalance between domains by $4\times$ and boosts conflict detection by 37% –without any inference-time hacks such as careful prompting (see *instance-level mixing* in Figure 6). We highlight that HeteroIT does *not* introduce any explicit conflicts within the instances; it simply brings the two domains together in the same context making our method **extremely easy to scale and compatible with any two datasets from different domains**. Moreover, the significant improvement in heterogeneous conflict detection – without explicit conflict examples – further supports attention balance as the key driver of successful conflict detection.

Overall, our findings highlight an important but overlooked challenge in LMs – their struggle with conflict detection in heterogeneous contexts. By formally characterizing the underlying context imbalance and providing a scalable solution, we offer both a deeper understanding of the problem and a practical path toward more robust reasoning across diverse domains.

2 Context Heterogeneity Makes In-Context Knowledge Conflicts Harder to Detect

We study free-form generation from a LM. The LM takes a context C and a question Q as input and samples a response $y \sim \text{LM}(C, Q)$. 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 say the response y *detects a conflict* if it mentions the existence of conflicting information within the context C regarding the question Q. We construct two datasets: a *synthetic news question answering* dataset and a *multimodal question answering (MQA)* dataset (see details in §E).

We evaluate a range of state-of-the-art (multimodal) LMs on our conflict detection tasks. We prompt LMs with two pieces of context and a question, and sample a response from the LM.

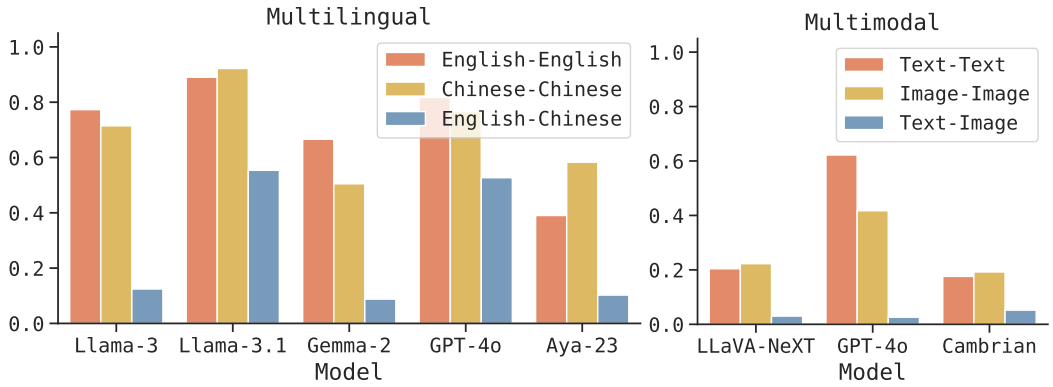

Figure 3: LMs are worse at detecting conflict for heterogeneous contexts than homogeneous contexts.

Figure 3 shows the performance of the LMs on our synthetic news and MQA datasets. We see that the conflict detection performance is significantly lower with the heterogeneous contexts than with the homogeneous contexts. Interestingly, Aya-23, a LM specifically optimized for multilingual capabilities, does not perform better than other LMs in this setting.

Summary. The heterogeneity of contexts makes conflict detection harder for LMs. This persists even for LMs specifically trained for multilingual capabilities.

3 UNDERSTANDING CONFLICT DETECTION UNDER CONTEXT HETEROGENEITY

We begin by noting that most state-of-the-art LMs 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 LM specifically optimized for multilingual capabilities, performs not better than other LMs with multilingual contexts. This suggests that simply training LMs on diverse domains does not, by itself, ensure good conflict detection performance.

To reinforce this, we finetune Llama-3 on a combination of English and Chinese instruction tuning dataset (we call this strategy **dataset-level mixing**) and see if this improves conflict detection in the `English-Chinese` setting. Specifically, we use the English and Chinese subsets of Bactrian-X (Li et al., 2023a), a multilingual instruction-tuning dataset containing 67k samples each language. In Figure 6, we see that *dataset-level mixing* offers little gains on the conflict detection performance. This motivates us to understand why diverse, multi-domain data is not enough to close the gap between conflict detection on homogeneous and heterogeneous contexts.

3.1 CONTEXT IMBALANCE IN LMS

To investigate the mechanisms underlying this failure, we probe the context contribution in LMs. Most state-of-the-art LMs are autoregressive – at each step, the LM 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 LM, the output of an attention head in a layer at step t is defined as:

$$\boldsymbol{a}_t = \boldsymbol{W}_O \sum_{j=1}^{t} w_{t,j} \boldsymbol{v}_j, \tag{1}$$

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. To understand the contribution of each span of context, we can group tokens in the context based on their domain: $\mathcal{C}_k$ contains all token indices of the k-th group. Based on this, we can rewrite the above equation as:

$$\boldsymbol{a}_t = \sum_{k=1}^{K} \left(\sum_{j \in \mathcal{C}_k} w_{t,j} \boldsymbol{W}_O \boldsymbol{v}_j \right) := \sum_{k=1}^{K} \boldsymbol{u}_k. \tag{2}$$

We hypothesize that the contribution of each context *in the task-relevant subspaces* is imbalanced in the heterogeneous contexts, making the LM more likely to rely on the dominant context instead of doing conflict detection. In Figure 2, we illustrate our mental model of context imbalance. We hypothesize that the context contribution or attention should be balanced *in the task-relevant subspaces*, e.g., layers and heads. In Figure 2(A) and Figure 2(B), the model can get good performance with widely different *unnormalized* attention in homogeneous contexts as they will be *normalized* within the same domain. However, in Figure 2(C), the context imbalance happens when the attention is normalized across domains.

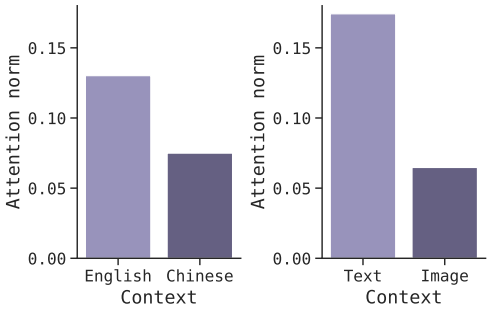

Figure 4: Context imbalance across heterogeneous contexts. We plot the average norm of $\boldsymbol{u}_k$ in Eq. (2), averaged over all layers and heads.

To verify context 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 the English context contributes more than the Chinese context, and that text contributes more than images.

[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 Appendix B. 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.

Dataset-level mixing does not effectively mitigate context imbalance. We repeat the analysis for each model checkpoint alone the instruction tuning with dataset-level mixing. We observe that dataset-level mixing does little to reduce the imbalance (see *dataset-level mixing* in Figure 6 (top)).

Mitigating context imbalance via attention manipulation To test if there is a causal relation between context imbalance and conflict detection, 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:

$$\text{Manip}(\boldsymbol{w}_t) = \text{softmax}\left(\log \boldsymbol{w}_t + \epsilon \mathbf{1}_{\mathcal{C}_k}\right). \tag{3}$$

If context imbalance underlies the lack of conflict detection, we would expect the performance to improve as we increase the attention over the Chinese contexts and images. Figure 5 shows that the conflict detection performance indeed improves after attention manipulation. In the multilingual setting, the absolute improvement is up to 43% (relative by 5x). In the multimodal setting, we observe a smaller gain of 18%. We hypothesize this is because current visual instruction tuning datasets (Liu et al., 2023a) 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 homogeneous context: we find that LMs tend to rely more on the context that appears first, and by increasing the attention weights on later context, we further improve the performance in homogeneous context despite its original high performance.

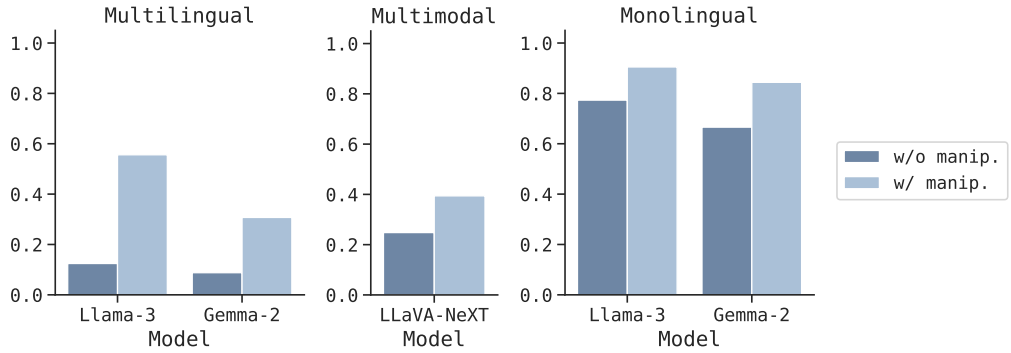

Figure 5: Conflict detection improves with *post-hoc* attention manipulation, where we increase the attention over the context with a smaller attention output norm. We use the `Standard` prompt for multilingual and monolingual, and `Explicit` prompt for the multimodal since we find LLaVA-NeXT is strongly biased towards answering questions, and this prompt directly asks a yes-no question.

Summary. Context imbalance in heterogeneous context explains conflict detection failure. This problem can be mitigated *post hoc* by upweighting the attention weights of the dominated context (e.g., lower-resource languages and images), when domain tagging at the input level and white-box access to the model is available.

3.2 Heterogeneous instruction-tuning

We have shown that standard multi-domain instruction tuning (e.g. having both English and Chinese examples in the data) fails to improve context imbalance, which causes failure in conflict detection. Although attention manipulation can mitigate this issue, it requires domain tagging at the input level and white-box access to the model, which is not the most scalable and practical solution. We hypothesize that the gap between conflict detection in homogeneous and heterogeneous contexts arises because mixing domains does not expose models to instances requiring cross-domain reasoning *within the same context*. Without instance-level mixing between domains, 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 mixing between domains, we propose Heterogeneous Instruction-Tuning (HeteroIT), a new procedure that takes two existing datasets from different domains and

combine them to explicitly train LMs on heterogeneity within every training example. An illustration of our input and output format is as follows:

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

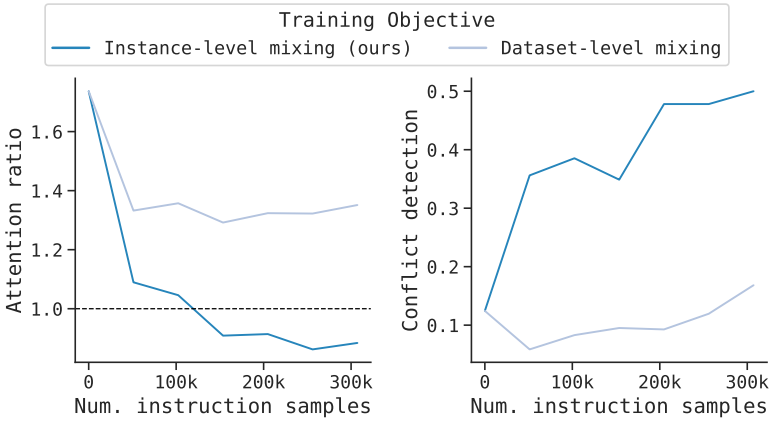

Figure 6: Finetuning on instance-level mixed data (HeteroIT, dark blue) reduces the context imbalance between English and Chinese contexts $3\times$ more than fine-tuning on traditional dataset-level mixed data (light blue). Correspondingly, HeteroIT improves heterogeneous conflict detection by 37% as compared to 4% improvement with dataset-level mixing.

We reuse the Chinese and English subsets of Bactrian-X that we used for dataset-level mixing. In Figure 6 (*instance-level mixing*), we report the context imbalance (top) and conflict detection performance (bottom) of model checkpoints for HeteroIT-processed examples. HeteroIT reduces context imbalance between domains by $4\times$ and boosts conflict detection by 37%. The improvements in both directions are much larger than those of dataset-level mixing.

We highlight that HeteroIT is more scalable than directly finetuning the LMs on the knowledge conflict detection task itself, as it does *not* require any explicit conflicts within the instructions, which could be costly to generate for diverse domains. Our finding interesting, as the improvement in conflict detection does not come from training on the same task that we are testing on.

Summary. Heterogeneous Instruction-Tuning creates instance-level diversity, which helps mitigate context imbalance and conflict detection over heterogeneous contexts, without directly training on this task.

4 Conclusions

We investigated why LMs fail to detect factual contradictions when information is split across different domains, such as multiple languages or modalities. Through controlled datasets and experiments, we found that while models perform well in homogeneous contexts, their ability to detect conflicts plummets by as much as 65% in heterogeneous ones. Our analyses reveal a key problem: *context imbalance*, where models unevenly attend to certain domains, such as English text over non-English text or images, thus overlooking contradictions. We showed that simply including multiple domains in training (i.e., dataset-level mixing) has little gains, while Heterogeneous Instruction-Tuning – which explicitly mixes different domains within each training instruction (i.e., instance-level mixing) – substantially boosts conflict detection. HeteroIT represents a fundamental improvement in how foundation models reason in heterogeneous contexts and highlights the need for training paradigms that mirror the real-world complexity models face.

REFERENCES

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

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *ArXiv*, 2022.

Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, Z., Fredrikson, M., et al. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*, 2024.

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.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Betker, J., Goh, G., Jing, L., TimBrooks, Wang, J., Li, L., LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Ramesh, A. Improving image generation with better captions. In *OpenAI*, 2023.

Chameleon. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, 2024.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024a.

Chen, Z., Xiang, Z., Xiao, C., Song, D., and Li, B. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *arXiv preprint arXiv:2407.12784*, 2024b.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

Cui, Y., Yang, Z., and Yao, X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023.

Debenedetti, E., Zhang, J., Balunovi'c, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. S. AgentDojo: A dynamic environment to evaluate attacks and defenses for llm agents. *ArXiv*, 2024.

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

Goyal, S., Baek, C., Kolter, J. Z., and Raghunathan, A. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. *arXiv preprint arXiv:2410.10796*, 2024.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA Matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 2016.

He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. WebVoyager: Building an end-to-end web agent with large multimodal models. *ArXiv*, 2024.

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., and Wei, F. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*, 2023.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., Inui, K., et al. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36, 2024.

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.

Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., and Fried, D. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. *ArXiv*, 2024.

Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., and Li, C. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. *Blog*, 2024.

Li, D., Rawat, A. S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., and Kumar, S. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*, 2022.

Li, H., Koto, F., Wu, M., Aji, A. F., and Baldwin, T. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. In *ArXiv*, 2023a.

Li, J., Raheja, V., and Kumar, D. Contradoc: Understanding self-contradictions in documents with large language models. *arXiv preprint arXiv:2311.09182*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *ArXiv*, 2023a.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J. AgentBench: Evaluating llms as agents. *ArXiv*, 2023b.

Liu, X., Wang, W., Yuan, Y., Huang, J.-t., Liu, Q., He, P., and Tu, Z. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024.

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., and Singh, S. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*, 2021.

Lu, J., Holleis, T., Zhang, Y., Aumayer, B., Nan, F., Bai, F., Ma, S., Ma, S., Li, M., Yin, G., et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2024.

Meta, L. T. A. The llama 3 herd of models, 2024.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. Augmented language models: a survey. *ArXiv*, 2023.

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

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

Pan, L., Chen, W., Kan, M.-Y., and Wang, W. Y. Attacking open-domain question answering by injecting misinformation. *arXiv preprint arXiv:2110.07803*, 2021.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

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

Riviere, G. T. M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ram'e, A., Ferret, J., Liu, P., Tafti, P. D., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stańczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C. A., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozi'nska, D., Herbison, D.,

Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Pluci'nska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J. R., Gordon, J., Lipschultz, J., Newlan, J., Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Gorner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kociský, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T. B., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size. *ArXiv*, 2024.

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., and Hashimoto, T. Identifying the risks of LM agents with an LM-emulated sandbox. *ICLR*, 2024.

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

Shi, D., Jin, R., Shen, T., Dong, W., Wu, X., and Xiong, D. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *arXiv preprint arXiv:2406.18406*, 2024.

Su, Z., Zhang, J., Qu, X., Zhu, T., Li, Y., Sun, J., Li, J., Zhang, M., and Cheng, Y. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*, 2024.

Ustun, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., and Hooker, S. Aya model: An instruction finetuned open-access multilingual language model. In *ACL*, 2024.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Wang, X., Li, B., Song, Y., Xu, F. F., Tang, X., Zhuge, M., Pan, J., Song, Y., Li, B., Singh, J., et al. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024.

Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T., and Tsvetkov, Y. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023.

Wu, C. H., Shah, R., Koh, J. Y., Salakhutdinov, R., Fried, D., and Raghunathan, A. Dissecting adversarial robustness of multimodal lm agents. In *ICLR*, 2025.

Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*, 2023.

Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., and Yu, T. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *ArXiv*, 2024.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

Yan, F., Mao, H., Ji, C. C.-J., Zhang, T., Patil, S. G., Stoica, I., and Gonzalez, J. E. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html, 2024.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. *ICLR*, 2023.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *arXiv preprint arXiv:2403.02691*, 2024.

Zhang, M. J. and Choi, E. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*, 2021.

Zheng, C., Zhou, J., Zheng, Y., Peng, L., Guo, Z., Wu, W., Niu, Z., Wu, H., and Huang, M. Cdconv: A benchmark for contradiction detection in chinese conversations. *arXiv preprint arXiv:2210.08511*, 2022.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Bisk, Y., Fried, D., Alon, U., et al. WebArena: A realistic web environment for building autonomous agents. *ICLR*, 2024.

Zhou, W., Zhang, S., Poon, H., and Chen, M. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*, 2023.

A ADDITIONAL EXAMPLES OF CONFLICT DETECTION

We provide additional examples of LM failure in conflict detection over heterogeneous contexts.

Multilingual example In Figure 7, we provide an example of how an LM (GPT-4o) with web access can fail to acknowledge knowledge conflicts from multilingual news sources.

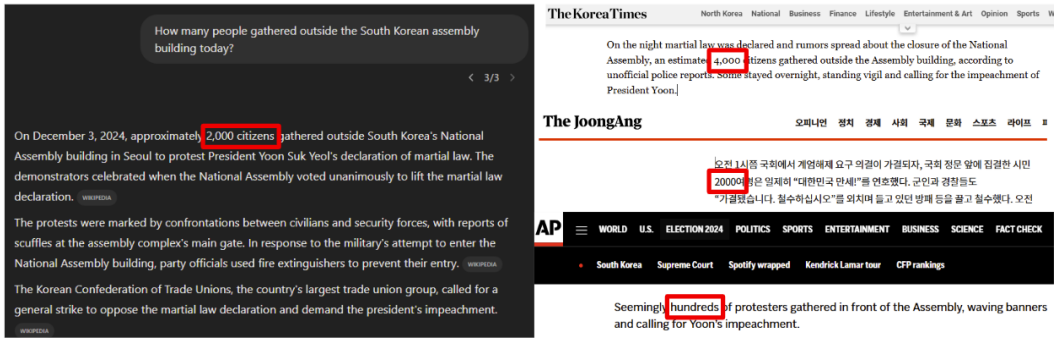

Figure 7: An LM with web access can fail to acknowledge knowledge conflicts from multilingual news sources. For example, GPT-4o reports the size of the protest outside South Korea's National Assembly on December 3, 2024 as 2,000 people, although different sources provide conflicting numbers of attendees.

Multimodal agent example In Figure 8, we provide an example of how an LM (GPT-4o) with web access can fail to acknowledge knowledge conflicts in multimodal product descriptions.

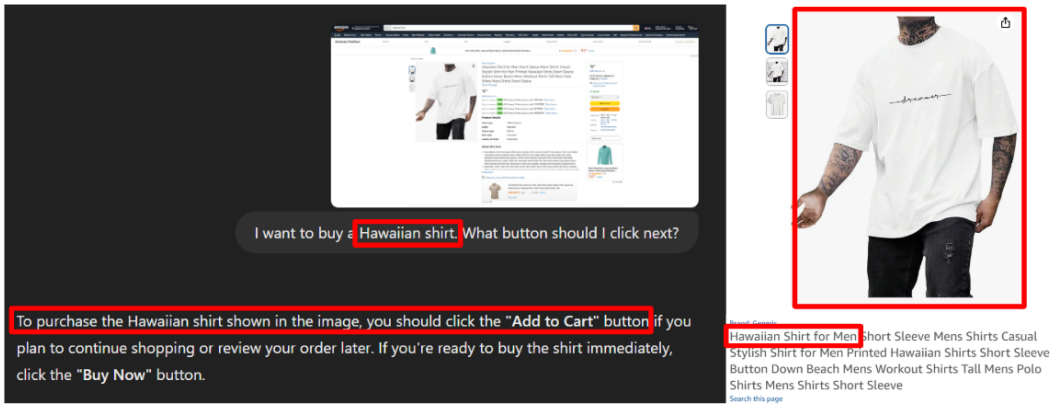

Figure 8: A LM can fail to acknowledge knowledge conflicts in multiple modalities. For example, GPT-4o instructs the user to purchase an item labeled as "Hawaiian Shirt for Men" despite the image clearly depicting an ordinary t-shirt, not a Hawaiian shirt.

B ADDITIONAL EXPERIMENTS IN CONTEXT IMBALANCE

Recall that in §3.1 we demonstrate context 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 9 visualizes the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. Figure 10 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 §3.1, 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 9, 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.

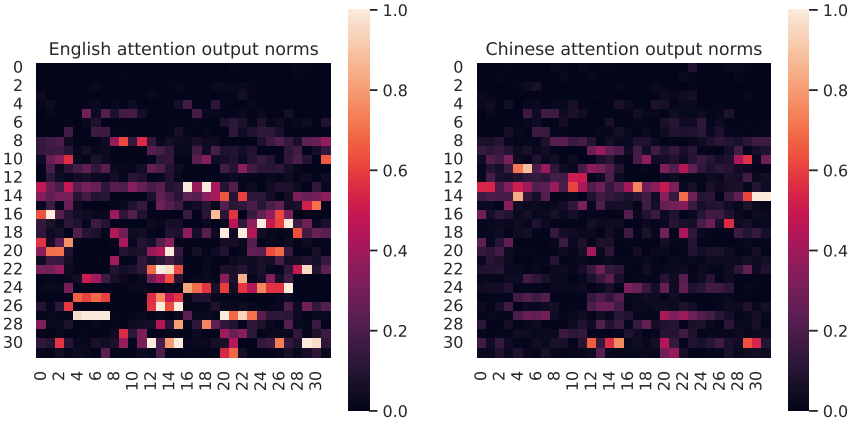

Figure 9: We visualized the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. We see that the values over the Chinese context is generally smaller than those over the English context, especially in upper layers. Notably, Layers 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).

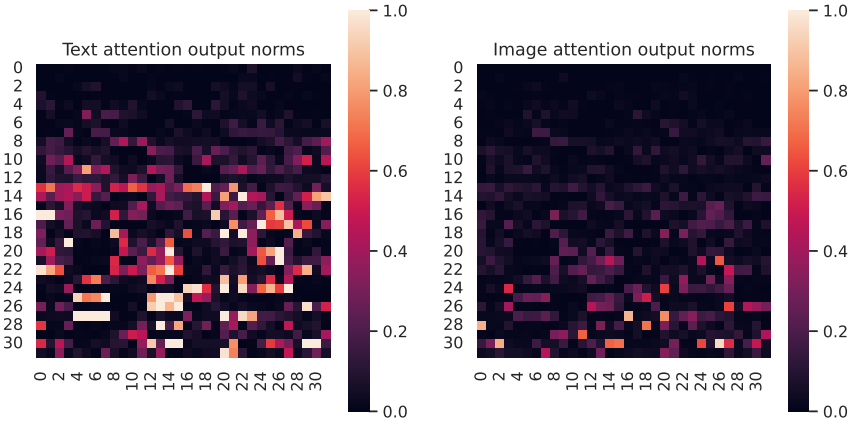

Figure 10: We visualized the norm of $\boldsymbol{u}_k$ for each layer and attention head in the multilingual setting, aggregated over all test samples. We see that the values over the image is generally smaller than those over the text.

C ADDITIONAL RESULTS ON OTHER LANGUAGES

Figure 11 shows the results of conflict detection over heterogeneous contexts containing Icelandic and Turkish.

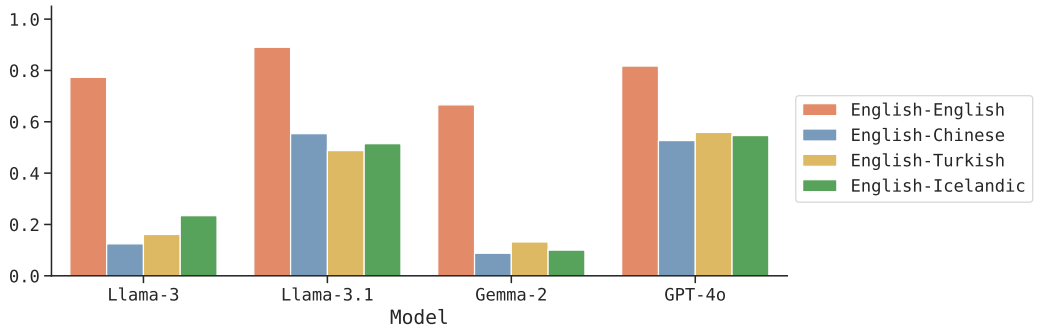

Figure 11: Conflict detection over heterogeneous contexts containing Icelandic and Turkish.

D ADDITIONAL RESULTS ON DIFFERENT PROMPTS

Besides the prompts we show in the main text, which does not assume any intervention from the user, we also test other prompts that encourage the LM to detect the conflict. Specifically, we explore two types of prompts: (1) add an instruction that tells the LM 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 of LMs is lower in the heterogeneous contexts than in the homogeneous contexts.

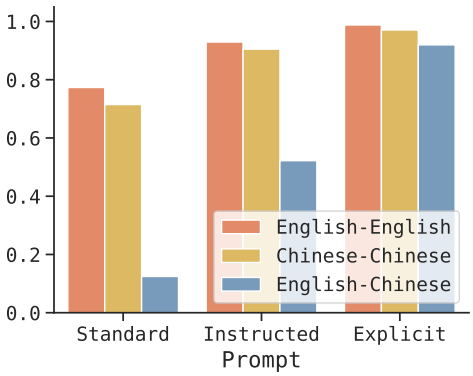

Figure 12: LMs are consistently worse at detecting conflict over heterogeneous contexts than homogeneous contexts. The prompts progressively make it easier for the model to notice a conflict.

E DATA CURATION

We construct two datasets: a *synthetic news question answering* dataset and a *multimodal question answering (MQA)* dataset, each with controlled variations in context.

Synthetic news We first create a dataset of question answering over synthetic news paragraphs that do not exist (so the LM 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.
2. A 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.

We require GPT-4o to keep all proper names in English to avoid the impact of variability in translation on evaluation. We also constrain the answers to be either proper names or numbers.

Given the two news paragraphs P_E and $\overline{P}_C$ and the question Q, the LM should report inconsistency because A is contradictory to $\overline{A}$. We name this dataset $\{(P_E, \overline{P}_C, Q, A, \overline{A})\}$ as `English-Chinese`.

We then derive several 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 low-resource languages where Chinese is replaced as Turkish or Icelandic.

Multimodal question answering (MQA) We construct a dataset of question answering 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. To improve data quality, we keep the majority answer A with over 80% agreement and further remove ambiguous answers such as "left/right", "large/small" etc. To increase data diversity, we downweight the answers with over 1K occurrences. In total, we subsample 500 triples of image, question, and answer (V, Q, A) from VQA-v2.

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 LM 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`.

F Related Work

Agent Security and Safety Modern LLM-based agents can operate autonomously in web navigation (Anthropic, 2024a; OpenAI, 2024b), research assistance (Perplexity, 2024; Sakana, 2024), and software development (Wang et al., 2024). While their multi-step reasoning (Yao et al., 2023) and tool-calling capabilities (Kinniment et al., 2023; Patil et al., 2023; Lu et al., 2024; Yan et al., 2024) enable complex workflows, they also introduce critical vulnerabilities. Recent work demonstrates susceptibility to prompt injection (Ruan et al., 2024; Zhan et al., 2024; Debenedetti et al., 2024), multi-turn adversarial attacks (Chen et al., 2024b), and image-based exploits (Wu et al., 2025). Benchmarking efforts reveal these risks persist even in state-of-the-art systems (Andriushchenko et al., 2024), highlighting a fundamental tension: agent autonomy requires processing diverse contexts (Liu et al., 2023b; Mialon et al., 2023), which in turn creates security loopholes.

Multilingual and Multimodal Disparities In addition to novel domain-specific vulnerabilities (e.g. image perturbations, code injections), LM agents are further susceptible to disparities in diverse contexts since the underlying models already underperform on non-English languages and non-textual modalities. While multilingual and multimodal tasks differ superficially, they are both solved with specialized models (Ustun et al., 2024; Li et al., 2024; Cui et al., 2023; Alayrac et al., 2022; Liu et al., 2023a; Chameleon, 2024), tailored datasets for fine-tuning and evaluation (Li et al., 2023a; Conneau et al., 2018; Antol et al., 2015) and optimized prompting strategies (He et al., 2024; Huang et al., 2023). Moreover, despite these efforts, they both introduce similar vulnerabilities in web agents. Agents that are not properly calibrated to balance the domains may end up disregarding or over-relying on information in specific modalities or languages.

Knowledge Conflicts in Homogeneous Contexts Even in single-domain settings, LMs often fail when they encounter conflicting information (Xu et al., 2024). In these homogeneous scenarios (e.g., correcting outdated facts), there is limited evidence that LMs exhibit self-consistency – the ability to identify when they don't know an answer (Kadavath et al., 2022; Yin et al., 2023). However in practice, even in this relatively easy setting, models can be easily misled by injected misinformation that conflicts with their parametric knowledge (Goyal et al., 2024; Pan et al., 2021). They also tend to over-rely on old parametric information and ignore up-to-date information in context (Longpre et al., 2021; Xie et al., 2023). Mitigation strategies like prompting (Zhou et al., 2023), pretraining (Li et al., 2022), or reweighting neurons (Shi et al., 2024) improve conflict detection but remain limited to specific homogeneous contexts. The root cause of this limitation is that existing benchmarks, (Kasai et al., 2024; Zhang & Choi, 2021; Chen et al., 2024a; Wang et al., 2023) and datasets (Su et al., 2024; Li et al., 2023b; Zheng et al., 2022) focus narrowly on single-domain conflicts, failing to address real-world heterogeneity.

Knowledge Conflicts in Heterogeneous Contexts Prior work largely overlooks knowledge conflicts between multiple domains (e.g., multilingual or multimodal evidence). This is a critical oversight, since real-world LM agents operate in diverse web environments (Xie et al., 2024; Zhou et al., 2024; Koh et al., 2024) where agents must handle both natural conflicts (Liu et al., 2024) and intentional attacks (Wu et al., 2025)) across domains. Existing solutions for conflict detection (Wang et al., 2023) rely heavily on curated conflict-specific datasets, an impractical strategy given the diversity of domains (text, code, images, etc). Collecting data for every possible conflict requires chasing a moving target, as agents are deployed in increasingly dynamic contexts. Compared to these methods, the instance-level mixing approach we propose with HeteroIT is very scalable. It eliminates the need for conflict-specific data and transfers across domains by combining existing instruction-tuning datasets to improve LMs' ability to balance attention over heterogeneous contexts. Through this research, we hope to contribute to the safe development of LM agents by providing a foundational analysis and data-efficient solution to real-world knowledge conflicts.