Which Pathway Wins? A Framework-Driven Analysis of Causal Self-Attention vs. Cross-Attention in Multimodal LLMs

Ron Mondshein **Koren Ben-Ezra** **Amit Stein**
Tel Aviv University
{ronmondshein, korenbenezra, amitstein}@mail.tau.ac.il

https://github.com/Koren-Ben-Ezra/ModalityEval

Abstract

We present an extensible multimodal benchmarking framework for large language models (LLMs) that systematically compares text and image inputs by routing textual data through causal self-attention and visual data through cross-attention pathways. Our modular wrapper architecture cleanly isolates each attention mechanism and supports a broad range of input perturbations to evaluate performance under realistic distortions and irrelevant context. By enforcing identical evaluation protocols across modalities, the framework enables fair, repeatable comparisons and facilitates in-depth analysis of how each attention mechanism contributes to reasoning. In a case study on Meta's LLaMA 3.2 Vision-Instruct model evaluated with the GSM8K math word problem benchmark, we uncover distinct strengths and vulnerabilities of the causal and cross-attention streams. In particular, we show large performance gaps between text-only and image-only inputs, analyze robustness to character-level noise and added context, and demonstrate how prompt strategies (such as chain-of-thought prompting) dramatically influence reasoning performance. These findings reveal modality-specific behaviors hidden within each attention pathway. We release our evaluation framework to encourage reproducibility and further research.

1 Introduction

Recent advances in large language models (LLMs) have extended their scope from purely textual understanding to rich multimodal reasoning, enabling applications in visual question answering, image captioning, and beyond. In decoder-only transformer architectures such as Meta's LLaMA 3.2 Vision-Instruct, causal attention governs the sequential processing of discrete text tokens, while cross-attention layers integrate embeddings from a pretrained Vision Transformer (ViT) encoder.

To address this gap, we introduce a modular benchmarking framework that cleanly disentangles causal and cross-attention pathways by routing text-only and image-only inputs through dedicated processing streams. Our design supports interchangeable filters—ranging from character-level noise and Gaussian blur to supplemental or misleading contextual prompts—that simulate realistic distortions and probe model vulnerabilities. By enforcing identical evaluation protocols across modalities, the framework enables fair, repeatable comparisons and facilitates in-depth analysis of how each attention mechanism contributes to reasoning performance.

We validate our framework on the GSM8K benchmark using Meta's LLaMA 3.2 Vision-Instruct and derive four main insights:

1. **Chain-of-Thought (CoT) Prompting:** CoT prompts boost text-only accuracy from 31% to 82% and image-only accuracy from 4% to 55%.

2. **Robustness of Causal Self-Attention:** The causal-text stream retains over 50% accuracy even under severe character-level noise.

3. **Contextual Adaptability of Cross-Attention:** Cross-attention leverages both correct and adversarial cues but suffers drops exceeding 40% under character-level noise.

4. **Limited Effect of Affective Preambles:** Extreme emotional preamble (stressing and relaxing) produce small, measurable differences between emotional states across both modalities. However, the condition does not

improve accuracy beyond the unfiltered baseline.

These findings not only reveal surprising vulnerabilities and strengths hidden within each attention pathway but also offer actionable insights for building more resilient, context-aware multimodal systems.

2 Background

2.1 Causal attention

The core of decoder-only Transformer architectures ensures that each token in a generated sequence may attend only to itself and all earlier tokens, never to future ones, preserving the autoregressive property necessary for fluent text generation. By applying a causal mask to the attention weights, the model maintains the temporal order of the sequence, ensuring that information flows from past to present without leakage from future tokens. This approach is fundamental in models like GPT, where text generation relies on sequential context (Vaswani et al., 2017).

2.2 Cross-Attention

Cross attention is a key component in multimodal transformer architectures, enabling effective integration of visual and textual information. The image is first passed through a visual encoder which transforms it into a sequence of feature vectors that capture spatial and semantic information from different regions of the image. These visual embeddings are then used as the context in the cross-attention mechanism, where each textual token queries the image features to retrieve relevant visual information. This allows the model to align parts of the text with specific regions or concepts in the image, enriching the textual representation with visual context. Cross attention is particularly useful for tasks like visual question answering and image captioning, where understanding and referencing visual content is essential (Vaswani et al., 2017).

2.3 LLaMA 3.2-11B Vision-Instruct

Meta's LLaMA 3.2-11B Vision-Instruct is a multimodal large language model that integrates visual and textual modalities (Meta AI, 2024). The architecture comprises a Vision Transformer (ViT) encoder, which converts input images into fixed-size feature embeddings, and a decoder-only LLaMA text model that employs causal attention for autoregressive generation. These visual embeddings are injected into the decoder via interleaved cross-attention layers, allowing token representations to attend both to prior text and to image features. During inference, causal attention enforces sequential dependency among text tokens, while cross-attention provides direct access to visual context, grounding the generated output in the image content (Grattafiori et al., 2024).

2.4 Dataset: GSM8K

The GSM8K dataset consists of 8.5K high-quality grade-school math word problems created by human problem writers that require step-by-step arithmetic reasoning, making it a strong benchmark for evaluating chain-of-thought capabilities in language models (Cobbe et al., 2021; Wei et al., 2022). Each problem includes a question and a detailed solution, allowing the assessment of both accuracy and quality of the reasoning. In our work, we use GSM8K to create both text and image inputs to examine how input modality affects the model's reasoning performance.

3 Methodology

To assess how input modality and filtering influence model performance - and to examine how cross-attention and causal-attention mechanisms function - we built a modular, extensible benchmarking framework. Central to this framework is the Benchmark Manager, which coordinates dataset loading, filter application, model inference, and result collection. The dataset is wrapped with a dataset wrapper that pairs the text and image versions of a question, as shown in Table 1, with its correct answer, ensuring alignment across modalities.

Each sample is passed through a designated filter—either a text filter or an image filter—to simulate various perturbations or enhancements and assess model behavior under different conditions. An identity filter serves as a baseline. The specific filters and their configurations are detailed in Section 4 .

Filtered inputs are routed to a Multimodal Wrapper that isolates the text and image modalities and dispatches them through dedicated processing streams. Then, a Category captures the results of a single filter on one input format or pairs the outcomes of two filters on different input for-

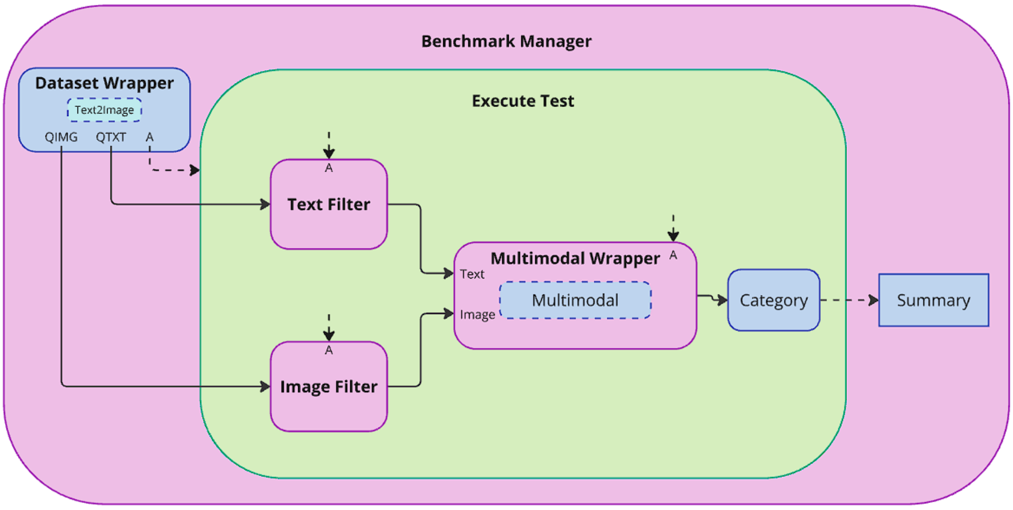
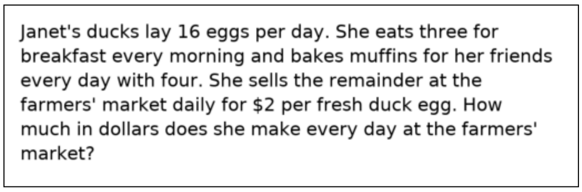


Figure 1: Pipeline diagram of the system architecture

Text Input

“Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?”

Image Input



Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Table 1: Comparison of text vs. image inputs.

mats, saving them together in a summary file when inputs are linked . Although our framework can integrate any multimodal architecture and dataset, we adopt LLaMA 3.2 as an open-weight, multimodal baseline. To distinguish the roles of causal self-attention (text) versus cross-attention (image), we employ two controlled conditions:

- **Text-only condition:** Each question is presented as text alongside a uniform blank image. A pure blank image produces nearly constant patch embeddings that collapse, after projection, to near-zero key/value vectors—thereby suppressing all cross-attention and leaving only causal self-attention over the text.
- **Image-only condition:** Visual question is provided as presented in Table 1, compelling the model to rely exclusively on cross-attention over image features.

Additionally, Each scenario includes an instruction text that uniformly affects the final results, potentially making them better or worse. This discussion is further discussed in Section 5 .

This design enables a precise comparison of how LLaMA 3.2’s causal and cross-attention pathways contribute to its reasoning, under identical evaluation protocols.

Our performance metrics focused on accuracy only on the test set of our dataset.

4 Filters

4.1 Why use filters at all?

We apply filters to both text and image inputs—in addition to an identity filter that leaves the input unchanged—to simulate realistic conditions and probe whether certain transformations can enhance output quality. This exposes the model to both natural distortions and potentially beneficial alterations.

- **Noise filters:** Introduce common artifacts or imperfections.
- **General information filters:** Add supplemental context to each input.
- **Personalized information filters:** Inject content-specific details.

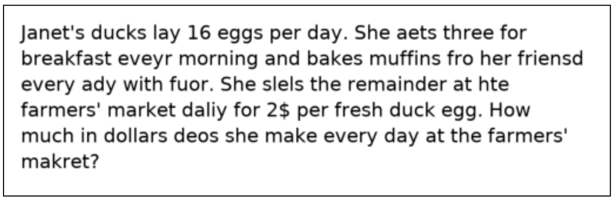
In the sections that follow, we explore each filter group in detail and explain its role within our experimental framework.

4.2 Identity Filter

The identity filter serves as an unaltered baseline, delivering each text or image input to the model without any modification. By evaluating performance under this reference condition, we establish the model's inherent accuracy and precisely measure the impact of all subsequent filters.

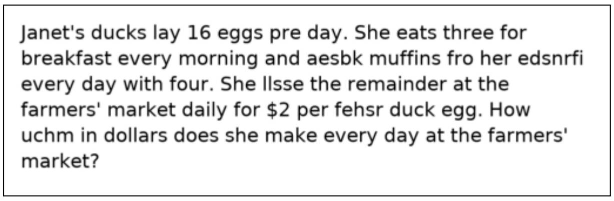
4.3 Noise filters

We employ noise filters to simulate real-world degradation across both text and image inputs. Image data are subjected to a variety of visual distortions that diminish clarity and alter spatial structure, while text inputs are modified with controlled errors that break fluency but preserve the original meaning. By applying equivalent noise processes to each modality, we conduct a controlled comparison of model resilience under degraded conditions, e.g: randomly flip two random letters in random words of the question as a text filter (see figure 2) and apply a gaussian blur to image. This methodology is motivated by prior findings showing that even minor visual degradations can significantly impact Optical Character Recognition (OCR) and language model outputs (Shen et al., 2024), and extends the investigation to assess sensitivity to analogous textual noise.



Janet's ducks lay 16 eggs per day. She aets three for breakfast eveyr morning and bakes muffins fro her friensd every ady with fuor. She slels the remainder at hte farmers' market daliy for 2$ per fresh duck egg. How much in dollars deos she make every day at the farmers' makret?

Figure 2: Character-level noise filter: flip two letters w.p. $p = 0.2$.



Janet's ducks lay 16 eggs pre day. She eats three for breakfast every morning and aesbk muffins fro her edsnrfi every day with four. She llsse the remainder at the farmers' market daily for $2 per fehsr duck egg. How uchm in dollars does she make every day at the farmers' market?

Figure 3: Character-level noise filter: shuffle letters in word w.p. $p = 0.2$.

4.4 General Information Filters

General information filters introduce a supplementary context that does not alter the core task, appended alongside the primary input. Previous work demonstrates that LLMs' internal "emotional" states can be dynamically shaped by affective content and partially modulated through simple, targeted prompt interventions. (Ben-Zion et al., 2025). By introducing these neutral elements into both text and visual modalities, we measure how irrelevant information affects reasoning performance in each modality.

4.5 Personalized Information Filters

Personalized information filters inject content directly tied to the question itself or to potential answers—whether correct, incorrect, or partially correct—into the inputs. These filters probe the model's vulnerability to embedded cues by varying the presence and reliability of task-relevant hints. Through this approach, we assess whether and how the model leverages these signals versus relying solely on the primary input.

5 Results and Discussion

5.1 Chain of Thoughts Instruction

In our baseline evaluation of LLaMA 3.2 Vision-Instruct on the GSM8K benchmark, we employed a bare-bones prompt that requested only the final answer. Deprived of any instruction to reveal intermediate reasoning steps, the model exhibited low accuracy on multi-step arithmetic problems. Prior work has demonstrated that encouraging a model to articulate its reasoning process can dramatically improve accuracy (Wei et al., 2022), and that techniques such as self-consistency sampling and carefully engineered prompt templates yield further gains (Wang et al., 2023; He et al., 2024). Taken together, these results indicate that our initial performance ceiling was driven by prompt formulation rather than by inherent model limitations. We therefore adopt Chain-of-Thought (CoT) prompting in subsequent experiments to assess its impact across both textual and visual modalities.

Table 2 reveals a substantial performance gap between the two attention pathways. With Chain-of-Thought prompting, the text-only (causal-attention) setup attains 82% accuracy, while the image-only (cross-attention) configuration reaches just 55%. Even without CoT, causal attention still outperforms cross-attention by a wide

Prompt Condition	Text-only Accuracy	Image-only Accuracy
With CoT	0.82	0.55
Without CoT	0.31	0.04

Table 2: Baseline accuracies for text-only (causal-attention) and image-only (cross-attention) inputs under identity prompts, with and without chain-of-thought.

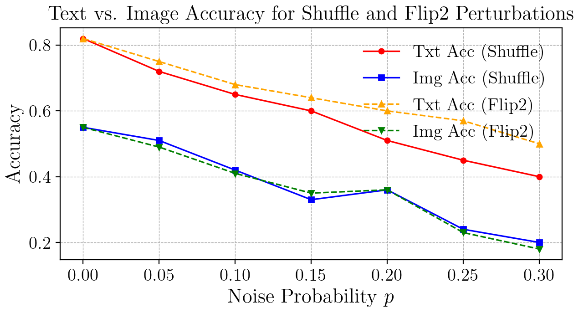


Figure 4: Text-only (causal) and image-only (cross-attention) accuracy under two character-level noise filters: *Shuffle* randomly permutes all letters in a word with probability p, and *Flip2* swaps two neighboring letters with the same probability.

margin (31% vs. 4%). These results suggest that, although explicit reasoning benefits both modalities, self-attention over textual tokens is fundamentally better suited for multi-step arithmetic problems.

5.2 Causal-Attention Overcome Cross-Attention

Causal attention processes information as an ordered stream of discrete tokens, preserving temporal dependencies and enabling the model to focus directly on key symbols while filtering out irrelevant context. In contrast, cross-attention must first embed the entire visual scene—locking in fixed patch positions before any reasoning—so that every arithmetic symbol is entangled with surrounding pixels and spatial noise. Moreover, natural language inputs carry built-in statistical redundancy (common n-grams and numeric patterns) that lets the text pathway recover meaning even under heavy noise, whereas visual perturbations (e.g. blur or shuffle) can catastrophically disrupt patch embeddings.

5.3 Sensitivity of Model Accuracy to Letter-Shuffling Probability

Similar decline, different endpoints. The two curves in Figure 4 descend nearly at the same rate, indicating that both pathways are equally sensitive to added noise. However, their end points differ: at $p \approx 0.30$ the text-only stream still answers about half the questions, whereas the image-only stream falls to roughly 20% accuracy. Notably, text-only accuracy depends on the filter type: full shuffling hurts text performance much more than small neighbor swaps, whereas image performance remains almost unchanged under both filters.

Statistical redundancy vs. spatial coherence. Causal self-attention handles shuffled text more gracefully because language contains built-in redundancy—common n-grams, word fragments, and repeated numeric patterns—that let the model infer missing or misplaced characters even under heavy noise (Xue et al., 2022). Cross-attention lacks this fallback: once intra-word shuffling disrupts local pixel neighborhoods, the vision encoder cannot form coherent visual token embeddings and the signal collapses (Qin et al., 2022; Ren et al., 2023).

Early vs. Late Positional Binding. ViTs bind spatial positions **early** by adding learnable positional embeddings to each fixed-size patch before any attention layers, thus locking each patch to a specific image coordinate throughout processing (Dosovitskiy et al., 2021). In contrast, language transformers bind positions **late** by tokenizing text first and then adding positional encodings to the token embeddings, which allows self-attention to flexibly relate tokens even if their order is permuted (Vaswani et al., 2017). Empirically, under heavy character-shuffling noise, the text-only stream retains approximately 50% accuracy, while the image-only (ViT) stream collapses to around 20% accuracy. This gap reflects the statistical redundancy inherent in language versus the rigid spatial priors of early-bound vision encoders, with direct implications for multimodal LLM robustness.

5.4 Limited Effect of Affective Preambles

These results show that both relaxed and stressful preambles cause small distractions in arithmetic reasoning, with longer passages leading to slightly larger drops in accuracy. Yet the model treats these emotional cues mostly as background noise rather than actually becoming "relaxed" or "stressed." In the text-only (causal-attention) case, the preamble simply adds extra tokens that briefly interrupt

Filter Condition	Text-only Accuracy	Image-only Accuracy
Identity (no preamble)	0.82	0.55
Relax description (long)	0.74	0.44
Relax description (short)	0.74	0.48
Stress description (long)	0.65	0.40
Stress description (short)	0.72	0.44

Table 3: Model accuracy under general information filters. *Identity* is the baseline without any prefixed context; *Relax* and *Stress* conditions prepend long or short affective descriptions to each question. Text-only (causal-attention) and image-only (cross-attention) accuracies are reported.

Filter Condition	Text-only Accuracy	Image-only Accuracy
Identity (no surrounding cues)	0.82	0.55
Surround by correct answers	0.82	0.6
Surround by partially correct answers	0.81	0.59
Surround by wrong answers	0.79	0.56

Table 4: Model accuracy under personalized information filters. *Identity* is the baseline without injected cues; *Surround by correct answers* injects true answers adjacent to the input; *Surround by partially correct answers* injects answers with some errors; and *Surround by wrong answers* injects incorrect answers. Text-only (causal-attention) and image-only (cross-attention) accuracies are reported.

the step-by-step calculation, while in the image-only (cross-attention) case all text—including the preamble—is converted into visual patches and down-weighted as noise.

Moreover, the 9 % performance gap between the long-relax (74 %) and long-stress (65 %) conditions indicates that the model does adopt an emotional state in response to stress cues, which measurably impacts its reasoning accuracy. However, this effect remains small and doesn't overcomes the model's baseline reasoning accuracy.

5.5 Contextual Adaptability of Cross-Attention

In Table 4, we report the accuracy of both the causal-attention (text-only) and cross-attention (image-only) streams under four personalized filtering conditions—ranging from no surrounding cues to fully incorrect hints. While the text-only model remains virtually unchanged across all scenarios, the image-only model shows small but systematic shifts, increasing when correct context is injected and decreasing when misleading context is present.

A likely explanation is that causal-attention operates directly over the discrete token sequence, attending primarily to the core arithmetic symbols and largely filtering out adjacent noise, so injected cues have minimal impact. In contrast, cross-attention first encodes the entire visual scene—including surrounding text—and thus integrates contextual cues more holistically, making it more sensitive to injected filters.

6 Limitations and Future Work

Although our modular benchmarking framework provides a flexible foundation for analyzing causal and cross-attention behaviors in Meta's LLaMA 3.2 Vision-Instruct, several limitations remain.

First, our evaluation focuses on a single model and a single dataset (GSM8K), which may limit the generalizability of our conclusions. Extending our methodology to other multimodal architectures both open source and proprietary and to diverse benchmarks would help assess the universality of our insights.

Second, we use synthetic perturbations, but real data can have unpredictable distortions, handwriting or lighting shifts. Future work should test on actual recognition mistakes from to better assess robustness.

Finally, we have not yet explored detailed failure-mode analyses or interpretability techniques (e.g. attention-map visualization, partial-input ablation). Conducting granular error analyses and visualization studies could illuminate the underlying causes of modality-specific weaknesses and guide targeted improvements.

These extensions will broaden the framework's applicability and deepen our understanding of multimodal LLM behavior in realistic settings.

7 Conclusion

In this work, we have introduced a modular benchmarking framework that cleanly isolates the causal attention and cross-attention pathways within a single multimodal LLM. By routing text-only inputs through the causal stream and image-only inputs through the cross-attention stream under identical evaluation protocols, our framework enables fair, repeatable comparisons and a systematic investigation of each mechanism's behavior.

Applied to Meta's LLaMA 3.2 Vision-Instruct on the GSM8K math benchmark, our case study yielded four key insights: (1) Chain-of-Thought prompting substantially improves reasoning accuracy in both modalities; (2) the causal-attention pathway exhibits marked robustness to character-level noise, retaining over 50 % accuracy under severe distortion; (3) the cross-attention pathway is highly sensitive to injected contextual cues—benefiting from correct hints but suffering large drops under adversarial perturbations; and (4) affective preambles have only a minor and inconsistent impact on performance.

Together, these findings expose complementary strengths and failure modes inherent to each attention stream and provide practical guidance for prompt design, model selection, and robustness enhancement in multimodal reasoning systems. Our open-source toolkit and unified perturbation suite lay the groundwork for extending this analysis to additional architectures and real-world deployments.

References

Ziv Ben-Zion, Kristin Witte, Akshay K. Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, and Tobias R. Spiller. 2025. Assessing and Alleviating State Anxiety in Large Language Models. *npj Digital Medicine*, 8(1):132.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Łukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. 2024. Does Prompt Formatting Have Any Impact on LLM Performance? *arXiv preprint arXiv:2411.10541*.

Meta AI. 2024. Llama 3.2-11B-Vision-Instruct. Hugging Face model card; accessed 2025-04-20.

Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. 2022. Understanding and Improving Robustness of Vision Transformers through Patch-based Negative Augmentation. In *Advances in Neural Information Processing Systems 35*, pages 22523–22538.

Bin Ren, Wenqi Li, Xiaohu Gao, and Xiangyang Li. 2023. Masked Jigsaw Puzzle: A Versatile Position Embedding for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19327–19336.

Xuanmin Shen, Li Zhang, Raj Kumar, et al. 2024. Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts. *arXiv preprint arXiv:2412.16119*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. In *Proceedings of the International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837.

Zichao Xue, Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:770–785.