Attention Under Attack: Analog Noise Effects and Mechanistic Vulnerabilities in Transformer Models

Mafizur Rahman and Lijun Qian
CREDIT Center, Department of ECE
Prairie View A&M University
Prairie View, Texas, USA
{mrahman13, liqian}@pvamu.edu

Abstract

Analog in-memory computing (AIMC) offers substantial efficiency gains for transformer inference but introduces hardware-induced noise that can distort attention behavior. Prior studies primarily focus on AIMC evaluations for vision tasks and CNN-based models. They largely overlook how hardware-induced noise perturbs internal attention dynamics in NLP models. In this work, we present the first fine-grained analysis of analog vulnerability in pretrained transformers, examining projection submodules, attention heads, and layer-wise dynamics across multiple NLP tasks. Results show that query (Q), key (K), and value (V) projections are the most sensitive components, while feed-forward layers remain comparatively robust. Also, analog noise yields depth-dependent degradation in higher layers, leading to scattered attention and disrupted token routing. This pre-deployment analysis mitigates potential resource misuse before physical deployment and offers practical guidance for designing noise-resilient analog NLP transformers.

1 Introduction

Natural language processing (NLP) research has increasingly explored efficiency-oriented deployment methods for transformer models, including quantization (Jin et al., 2024), low-precision inference (Zadeh et al., 2020), and pruning (Ilhan et al., 2024), with a primary focus on preserving task accuracy under reduced digital precision. However, as model sizes continue to increase, the computational and energy demands of transformer inference have grown substantially. This growth has created strong interest in analog in-memory computing (AIMC) as a more energy-efficient alternative to traditional digital accelerators. Recent work in vision (Lambertini et al., 2025) and speech demonstrates that AIMC hardware can significantly improve energy efficiency and latency (Zadeh et al., 2020). However, state of the art (SOTA) AIMC simulators (Lammie et al., 2022; Xiao et al., 2022; Chen et al., 2018) primarily support CNN-based models and do not fully support transformer architectures, with AIHWKIT (Rasch et al., 2021) being a notable exception. Thus, the impact of AIMC hardware on language models remains unclear. Unlike vision or speech models, NLP systems rely on precise attention patterns and deep contextual reasoning, which may respond very differently to analog noise.

A central challenge is that analog hardware introduces multiple sources of non-idealities such as quantization limits, device drift, read noise, and stochastic cell behavior, which can perturb internal representations (Brooks et al., 2020). Deploying transformer models directly onto analog hardware without prior robustness analysis risks severe or unpredictable performance degradation (Ke et al., 2025; Latibari et al., 2024). At the same time, access to large-scale analog hardware platforms is extremely limited. Consequently, practitioners lack guidance on which transformer models, internal components, and attention mechanisms are most vulnerable when mapped to analog accelerators.

To address this gap, we conduct a systematic pre-deployment robustness study of transformer models using AIHWKIT (Rasch et al., 2021), a hardware-aware simulator calibrated to PCM-style analog devices. Simulation is not a replacement for real hardware; rather, it enables controlled and repeatable experiments across hardware fidelity settings that are impractical or impossible on physical devices due to cost, throughput, and limited availability. Therefore, our goal is not to replicate exact chip behavior, but to provide actionable design insights for researchers and hardware practitioners. Overall, our analysis is guided by two research questions: **RQ1: Benchmarking Analog Inference Robustness:** *How does analog hardware noise affect end-to-end performance across*

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 227–237

transformer architectures and NLP tasks?

RQ2: Mechanistic Vulnerability: *Which internal components-submodules, layers, and attention heads—are most sensitive to analog perturbations, and how does analog noise reshape attention behavior?*

By combining task-level evaluation with submodule-level isolation and attention-pattern analysis, our work provides a comprehensive view of **where**, **how**, and **why** analog noise disrupts transformer reasoning in NLP. Our findings offer a practical pre-deployment roadmap for AIMC researchers and hardware designers. Also, it enables more robust analog transformer designs before committing to expensive fabrication or device-level experimentation. Related work on NLP and AIMC hardware is provided in **Appendix A.4**.

2 Analog Inference Framework

Models and Datasets. We study five representative transformer families: BERT-base (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), RoBERTa-base (Liu et al., 2019), and DeBERTa-large (He et al., 2020) across three standard NLP tasks: Sentiment Classification (SST-2 (Socher et al., 2013)), Natural Language Inference (MNLI (Williams et al., 2018)), and Question Answering (SQuAD v1.1 (Rajpurkar et al., 2016)). To extend to decoder LLMs, we evaluate Phi-3-mini and LLaMA-3.2-1B (8-bit) on ARC-Easy (Clark et al., 2018) (dataset details in **Appendix A.6**).

Analog Hardware Simulation. We simulate analog transformer inference using AIHWKIT, IBM's open-source hardware-aware emulator calibrated to PCM and ReRAM devices. Analog inference proceeds by first loading pretrained transformer models and NLP datasets in the digital domain. Later, we transform all linear layers within each transformer block into analog-compatible operators (e.g., `AnalogLinear`). During execution, digital input activations are quantized and converted to analog signals via digital to analog converters (DACs), then applied to analog crossbar arrays where weights are mapped as conductance values (see **Appendix A.3** for details on mapping transformer weights onto analog crossbar arrays). The crossbars perform noisy analog matrix-vector multiplications (MVMs) in parallel, incorporating device-level non-idealities such as programming noise, read noise, drift, and finite ADC/DAC resolution as modeled by AIHWKIT. Partial analog currents are integrated and digitized through ADCs, summed across crossbar tiles when needed, and passed to subsequent digital components. This process repeats layer by layer, allowing us to observe how analog noise introduced during MVM propagates through attention and feed-forward blocks to influence the final model output. An overview of the analog transformer inference pipeline is illustrated in Figure 1.

Since AIHWKIT does not expose explicit cell-level precision, we simulate analog hardware fidelity by jointly controlling multiple sources of non-idealities that arise during analog computation. Analog MVM is modeled as

$$\mathbf{y} = Q_{\mathrm{ADC}}((\mathbf{W} + \Delta\mathbf{W}) \cdot Q_{\mathrm{DAC}}(\mathbf{x}) + \boldsymbol{\epsilon}_{\mathrm{out}}), \quad (1)$$

where $\mathbf{x}$ denotes the digital input vector, $\mathbf{W}$ the programmed conductance matrix, $\Delta\mathbf{W}$ device-level perturbations, $Q_{\mathrm{DAC}}(\cdot)$ and $Q_{\mathrm{ADC}}(\cdot)$ finite-resolution digital-to-analog and analog-to-digital quantization, and $\boldsymbol{\epsilon}_{\mathrm{out}}$ output noise.

We introduce analog non-idealities through five coupled mechanisms: finite ADC/DAC resolution, stochastic weight-modifier noise during programming, read noise during inference, output noise after accumulation, and PCM-specific programming, read, and drift noise modeled via a calibrated noise process. Bias terms are retained in digital form, and learnable output scaling is enabled to stabilize analog activations. Lastly, global drift compensation is used to mitigate long-term conductance drift. All experiments are conducted under a fixed noise configuration to reflect realistic operating conditions without entering overly optimistic or catastrophic regimes. Models are evaluated in inference-only mode without fine-tuning. To concretely illustrate the analog transformation, consider a standard self-attention projection in a transformer layer. In the digital model, the query projection is computed as

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q + \mathbf{b}_Q, \quad (2)$$

where $\mathbf{X} \in \mathbb{R}^{T\times d}$ denotes token embeddings, $\mathbf{W}_Q \in \mathbb{R}^{d\times d}$ the learned projection matrix, and $\mathbf{b}_Q$ the bias term.

Under analog execution, $\mathbf{W}_Q$ is mapped to a conductance matrix stored in the crossbar array, and the projection is computed as

$$\mathbf{Q}^{\mathrm{ana}} = Q_{\mathrm{ADC}}((\mathbf{W}_Q + \Delta\mathbf{W}_Q) \cdot Q_{\mathrm{DAC}}(\mathbf{X}) + \boldsymbol{\epsilon}) + \mathbf{b}_Q, \quad (3)$$

where $\Delta\mathbf{W}_Q$ captures device-level noise and drift, and $\boldsymbol{\epsilon}$ aggregates read and output noise. The same

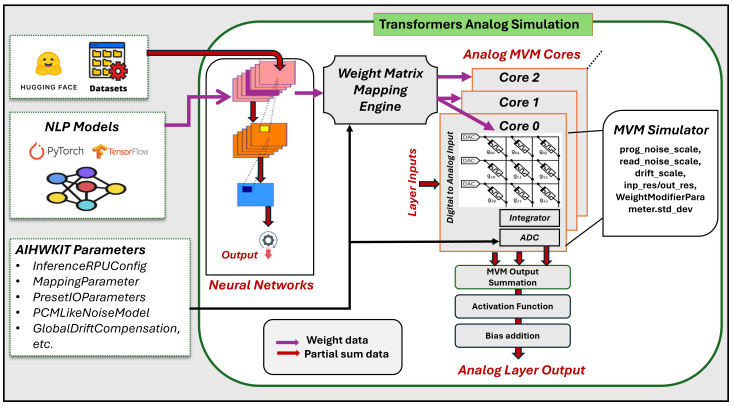

Figure 1: End-to-end workflow of analog transformer inference using AIHWKIT. Pretrained NLP models are mapped to analog crossbar arrays, where linear layers execute noisy analog MVMs with device non-idealities.

transformation is used independently to key, value, attention output, and feed-forward layers. We discuss the practical challenges of transformer layer conversion in **Appendix A.1**.

Submodule-Level Noise Isolation. We perform targeted submodule sensitivity analysis to specify which internal components are most sensitive to analog noise. For each experiment, the full digital model is first converted to analog execution using the medium-noise configuration. Subsequently, all analog linear layers except a selected target submodule are replaced with ideal (noise-free) analog equivalents that preserve pretrained weights. This selective replacement ensures that only the target submodule experiences analog non-idealities, while all other components behave identically to their digital counterparts. We evaluate noise sensitivity for query, key, and value projection layers, the attention output projection, and the FFN intermediate and output layers. Performance degradation relative to the fully digital baseline directly reflects the contribution of each submodule to overall analog robustness.

Attention Head and Layer Vulnerability. We analyze how analog noise alters attention behavior itself beyond submodule-level effects. For each attention head, we compare digital and analog attention distributions over input tokens. Distributional deviation is quantified using the Kullback–Leibler (KL) divergence (Hershey and Olsen, 2007)

$$\mathrm{KL}\left(A^{\mathrm{dig}} \,\|\, A^{\mathrm{ana}}\right) = \sum_i A_i^{\mathrm{dig}} \log \frac{A_i^{\mathrm{dig}}}{A_i^{\mathrm{ana}}}, \tag{4}$$

where A^{dig} and A^{ana} represent normalized attention weights produced by the digital and analog models. To characterize qualitative changes in attention structure, we compute the entropy shift

$$\Delta H = H(A^{\mathrm{ana}}) - H(A^{\mathrm{dig}}), \quad H(A) = -\sum_i A_i \log A_i. \tag{5}$$

Negative ΔH indicates attention focusing, while positive ΔH corresponds to attention scattering. Aggregating entropy shifts and divergence scores across heads and layers reveals depth-dependent vulnerability patterns and identifies which semantic stages of transformer reasoning are most disrupted by analog noise.

Experiment Setup: Analog mapping parameters include tile size = 256, column-wise scaling, and fixed weight clipping with bound 1.0. The noise configuration uses a weight modifier standard deviation of 0.05, read noise of 0.0175, output noise of 0.04, and PCM programming, read, and drift noise scaling factors set to 1.0 with global drift compensation. Results are averaged over three independent runs (random seeds = 0, 1, 2). All simulations are conducted using AIHWKIT (v0.9.2) under a consistent inference-only setting. More details are in **Appendix A.2**.

3 Results

RQ1: Benchmarking Analog Inference Robustness. Table 1 reports end-to-end digital versus analog inference performance across MNLI, SST-2, and SQuAD under a fixed analog noise configuration. Across all tasks and models, analog inference presents a measurable but non-catastrophic performance degradation. Accuracy and F1 drops remain within 1–3 points for most encoder-based models, suggesting that pretrained transformers retain substantial functional robustness under realis-

Task	Model	Digital	Analog	Drop Δ
MNLI	DistilBERT	$80_{\pm 0.2}$ / $0.99_{\pm 0.01}$	$79_{\pm 0.4}$ / $0.98_{\pm 0.02}$	−1.0 / −0.01
	BERT	$84_{\pm 0.3}$ / $0.92_{\pm 0.01}$	$83_{\pm 0.5}$ / $0.94_{\pm 0.02}$	−1.0 / +0.02
	RoBERTa	$87_{\pm 0.3}$ / $0.85_{\pm 0.01}$	$85_{\pm 0.6}$ / $0.88_{\pm 0.02}$	−2.0 / +0.03
	DeBERTa	$91_{\pm 0.3}$ / $0.48_{\pm 0.02}$	$89_{\pm 0.6}$ / $1.40_{\pm 0.05}$	−2.0 / +0.92
SST-2	BERT	$93_{\pm 0.2}$ / $0.24_{\pm 0.01}$	$92_{\pm 0.3}$ / $0.25_{\pm 0.01}$	−1.0 / +0.01
	ALBERT	$93_{\pm 0.3}$ / $0.21_{\pm 0.01}$	$90_{\pm 0.5}$ / $0.26_{\pm 0.02}$	−3.0 / +0.05
	RoBERTa	$94_{\pm 0.2}$ / $0.22_{\pm 0.01}$	$92_{\pm 0.4}$ / $0.23_{\pm 0.01}$	−2.0 / +0.01
SQuAD	BERT	$89_{\pm 0.3}$ / $87_{\pm 0.4}$	$88_{\pm 0.5}$ / $85_{\pm 0.6}$	−1.0 / −2.0
	RoBERTa	$95_{\pm 0.2}$ / $93_{\pm 0.3}$	$89_{\pm 0.6}$ / $87_{\pm 0.5}$	−6.0 / −6.0
	DeBERTa	$95_{\pm 0.3}$ / $92_{\pm 0.4}$	$92_{\pm 0.4}$ / $90_{\pm 0.5}$	−3.0 / −2.0

Table 1: **Digital vs. analog inference performance across NLP tasks.** MNLI and SST-2 report Acc/Loss; SQuAD reports F1/EM. Δ denotes performance drop (-) or loss increase (+) under analog inference.

tic analog noise. However, task difficulty strongly modulates analog sensitivity. Based on Table 1 results, classification tasks, such as MNLI and SST-2, exhibit limited degradation. BERT and DistilBERT show approximately one-point accuracy drop, while RoBERTa and ALBERT experience slightly larger drops. In contrast, SQuAD exhibits noticeably higher sensitivity. Both RoBERTa and DeBERTa incur 3–6 point reductions in both F1 and exact match. This gap reflects the reliance of question answering on precise token-level alignment and sharp attention distributions. These results demonstrate that analog noise does not uniformly affect all NLP tasks. Instead, its impact grows with the need for fine-grained contextual reasoning. Overall, the global analysis establishes that analog transformer inference is feasible for NLP. It also reveals systematic task- and model-dependent robustness differences that encourage deeper investigation into submodules and attention levels in the following sections.

Extension to Decoder-Style LLMs. Additional experiments are conducted on decoder-style LLMs, specifically Phi-3-mini and LLaMA-3.2-1B (8-bit), using AIHWKIT. Zero-shot performance is evaluated on the ARC-Easy benchmark under three analog noise regimes (low, medium, high). Performance degradation scales consistently with noise intensity, with modest drops of approximately 2–3% under medium noise and larger degradation under high noise (up to ∼7%). These trends are consistent with the encoder-based results reported in Table 1, indicating that analog robustness characteristics generalize across model families. Despite architectural discrepancies, the qualitative behavior remains stable: increasing analog noise primarily affects reasoning accuracy in a smooth, predictable manner, without catastrophic failure. All experiments use identical preprocessing, sequence length (2048), and no noise-aware retraining, guaranteeing a consistent comparison across noise regimes.

Noise	modifier_std	out_noise	Phi-3 Δ	LLaMA-3.2-1B Δ
Low (0.5×)	0.025	0.02	−1.3	−1.4
Medium (1×)	0.05	0.04	−2.6	−2.6
High (2×)	0.10	0.08	−7.1	−6.4

Table 2: **Analog noise sensitivity for decoder-style LLMs on ARC-Easy (zero-shot).** Performance degradation (Δ) is reported across three noise regimes.

RQ2: Mechanistic Vulnerability. We find that analog noise impacts transformer submodules in a highly non-uniform manner. Across all tasks and architectures, Q, K, and V projection layers are the most vulnerable, causing the largest performance drops when isolated under analog execution (see Figure 2). The vulnerability of Q, K and V projections is consistent across architectures but differs in magnitude due to architectural design choices. In contrast, feed-forward network (FFN) layers show comparatively higher robustness, with smaller degradation across SST-2 and MNLI, indicating that FFNs can partially absorb analog perturbations. Interestingly, models with heavier reliance on disentangled or reweighted attention representations, such as DeBERTa, show disproportionately larger degradation when these projections are perturbed, even under moderate analog noise. This indicates that analog perturbations introduced at the attention formation stage distort token–token relevance scores in a way that downstream layers cannot fully correct. In contrast, FFN layers operate on already aggregated contextual representations and therefore experience a reduced impact of noise, specifically for sentence-level classification tasks. Attention output projections exhibit intermediate robustness, being more stable than Q, K and V, but still contribute to performance loss in reasoning-intensive tasks, such as SQuAD. These trends suggest that analog noise primarily disrupts *where* the model attends, rather than *how* it transforms the attended information. Overall, analog robustness is submodule-dependent with attention projections dominating failure behavior.

Attention-Level Vulnerability and Behavioral Shifts. The attention-level analysis shows that analog noise introduces structured and non-uniform degradation across layers and heads. Head sensitiv-

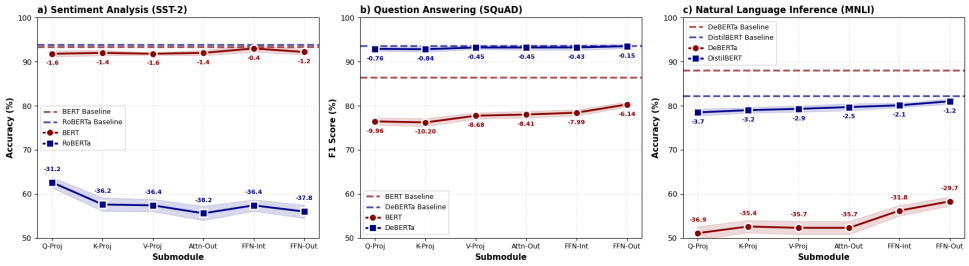

Figure 2: Submodule analysis of transformer architectures across three NLP tasks under analog noise. Bold solid lines denote analog inference results, while dashed lines indicate digital baselines. Negative values indicate degradation due to noise.

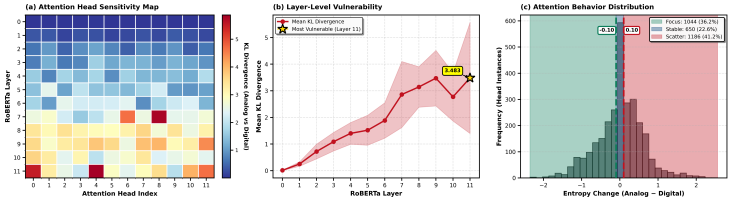

Figure 3: Layer- and head-level attention vulnerability under analog noise for the RoBERTa question answering.

ity maps (Figure 3a) show that vulnerability concentrates in higher RoBERTa layers with a small subset of heads exhibiting consistently higher divergence while many remain relatively stable. This trend is reinforced by the layer-wise KL divergence, which grows monotonically with depth and peaks in the final layers (Figure 3b). This suggests that higher-level semantic reasoning is more sensitive to analog perturbations. Lastly, beyond magnitude alone, attention behavior analysis reveals a clear shift toward increased scattering with a larger fraction of heads exhibiting a positive entropy change compared to focused or stable regimes (Figure 3c). This redistribution implies that analog noise primarily disrupts precise token routing rather than uniformly corrupting attention, leading to diffuse attention patterns that degrade reasoning-intensive performance.

We also discuss input token length impact and theoretical energy scaling in **Appendix A.5**.

Benefits of AIMC and Robustness Considerations. Prior AIMC systems demonstrate significant practical benefits, including improved energy efficiency (e.g., up to ∼658 TOPS/W in SRAM-based designs) and competitive accuracy under realistic device constraints (Sebastian et al., 2020; Rasch et al., 2024). However, these demonstrations have mainly focused on CNN-based models for image classification tasks. Additionally, unlike digital quantization methods, which introduce deterministic and independent rounding noise that remains fixed at inference time, AIMC introduces structured and time-varying perturbations arising from device-level effects such as programming noise, conductance drift, and read variability. These perturbations directly affect MVMs and can induce correlated shifts in attention computations, leading to systematic changes in attention patterns rather than uniform degradation. *Our mechanistic analysis addresses this discrepancy by showing that such structured analog perturbations are not captured by conventional digital approximation methods.*

4 Conclusion

In this study, using hardware-aware simulation, we show that analog noise induces consistent but task- and architecture-dependent performance degradation, with larger impact on reasoning-intensive tasks than on sentence-level classification. Submodule- and attention-level analyses show that Q/K/V projections and early attention formation are the most vulnerable, while feed-forward layers remain relatively robust. Also, analog inference does not fail uniformly but disrupts specific internal mechanisms that govern attention structure and contextual reasoning. These insights offer practical guidance for designing more robust analog transformers and motivate future work on hardware-aware attention stabilization.

Acknowledgment: This research is supported by the Army Research Office under Cooperative Agreement W911NF-24-2-0133.

5 Limitations

One limitation of this study is that it relies on hardware-aware simulation rather than deployment on fabricated analog accelerators. However, AIMC simulators such as AIHWKIT have been extensively validated in prior work and provide a reliable approximation of device non-idealities, noise, and variability for pre-deployment evaluation. As a result, our findings should be interpreted as indicative robustness trends rather than exact predictors of absolute on-chip performance. Another limitation is that we focus exclusively on AIHWKIT simulator because it currently provides the most reliable and extensible support for transformer architectures, whereas other SOTA analog simulators remain limited to CNNs or shallow MLPs and cannot faithfully model modern attention-based systems.

Finally, we do not report absolute energy measurements, as our primary goal is robustness and mechanistic analysis. Accurate end-to-end energy estimation for analog transformer inference remains infeasible without deployable hardware and standardized power measurement interfaces.

6 Ethical Considerations

Our work does not present new datasets, personal data, or downstream decision-making systems and therefore poses minimal direct ethical risk. However, deploying transformer models on analog hardware without adequate robustness analysis could lead to silent performance degradation, particularly in high-stakes applications such as medical question answering or legal document analysis. Therefore, by explicitly identifying vulnerable submodules, layers, and attention mechanisms, this study aims to reduce such risks rather than exacerbate them. We highlight that analog deployment should be preceded by thorough validation and transparency about potential failure modes.

References

Stefano Ambrogio, Pritish Narayanan, Atsuya Okazaki, Andrea Fasoli, Charles Mackin, Kohji Hosokawa, Akiyo Nomura, Takeo Yasuda, An Chen, A Friz, and 1 others. 2023. An analog-ai chip for energy-efficient speech recognition and transcription. *Nature*, 620(7975):768–775.

David Brooks, Martin M Frank, Tayfun Gokmen, Udit Gupta, X Sharon Hu, Shubham Jain, Ann Franchesca Laguna, Michael Niemier, Ian O'Connor, Anand Raghunathan, and 1 others. 2020. Emerging neural workloads and their impact on hardware. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1462–1471. IEEE.

Julian Büchel, William Andrew Simon, Corey Lammie, Giovanni Acampa, Kaoutar El Maghraoui, Manuel Le Gallo, and Abu Sebastian. 2024. Aihwkit-lightning: a scalable hw-aware training toolkit for analog in-memory computing. In *NeurIPS 2024 Workshop Machine Learning with new Compute Paradigms*.

Pai-Yu Chen, Xiaochen Peng, and Shimeng Yu. 2018. Neurosim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(12):3067–3080.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John R Hershey and Peder A Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Fatih Ilhan, Gong Su, Selim Furkan Tekin, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Resource-efficient transformer pruning for finetuning of large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16206–16215.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12186–12215.

Vinay Joshi, Manuel Le Gallo, Simon Haefeli, Irem Boybat, Sasidharan Rajalekshmi Nandakumar, Christophe Piveteau, Martino Dazzi, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. 2020. Accurate deep neural network inference using computational phase-change memory. *Nature communications*, 11(1):2473.

Jia Ke, Wang Xiaohao, Chen Hailin, Zhong Wei, Li Xinxiong, Fang Zenan, and An Fengwei. 2025. A component-centric perspective on hardware accelerators for llms. *IEEE Access*.

Alessandro Lambertini, Tommaso Zanotti, Paolo Pavan, Andrea Padovani, and Francesco Maria Puglisi. 2025. Simulation and benchmarking of crossbar parasitic resistance models: Accuracy and performance comparison. *APL Machine Learning*, 3(2).

Corey Lammie, Wei Xiang, Bernabé Linares-Barranco, and Mostafa Rahimi Azghadi. 2022. Memtorch: An open-source simulation framework for memristive deep learning systems. *Neurocomputing*, 485:124–133.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Banafsheh Saber Latibari, Najmeh Nazari, Muhtasim Alam Chowdhury, Kevin Immanuel Gubbi, Chongzhou Fang, Sujan Ghimire, Elahe Hosseini, Hossein Sayadi, Houman Homayoun, Soheil Salehi, and 1 others. 2024. Transformers: A security perspective. *IEEE Access*.

Manuel Le Gallo, Corey Lammie, Julian Büchel, Fabio Carta, Omobayode Fagbohungbe, Charles Mackin, Hsinyu Tsai, Vijay Narayanan, Abu Sebastian, Kaoutar El Maghraoui, and 1 others. 2023. Using the ibm analog in-memory hardware acceleration kit for neural network training and inference. *APL Machine Learning*, 1(4).

Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuan-Jing Huang, and Xipeng Qiu. 2022. Towards efficient nlp: A standard evaluation and a strong baseline. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3288–3303.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mafizur Rahman, Lin Li, Lijun Qian, and Max Huang. 2024. Comparative analysis of inference performance of pre-trained deep neural networks in analog accelerators. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 468–475. IEEE.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Malte J Rasch, Fabio Carta, Omobayode Fagbohungbe, and Tayfun Gokmen. 2024. Fast and robust analog in-memory deep neural network training. *Nature Communications*, 15(1):7133.

Malte J Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. 2021. A flexible and fast pytorch toolkit for simulating training and inference on analog crossbar arrays. In *2021 IEEE 3rd international conference on artificial intelligence circuits and systems (AICAS)*, pages 1–4. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. 2020. Memory devices and applications for in-memory computing. *Nature nanotechnology*, 15(7):529–544.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty Van Aken, Qingqing Cao, Manuel R Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, and 1 others. 2023. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122.

T Patrick Xiao, Christopher H Bennett, Ben Feinberg, Matthew J Marinella, and Sapan Agarwal. 2022. Crosssim: accuracy simulation of analog in-memory computing. *GitHub, v2. 0, https://github.com/sandialabs/cross-sim*, page 2.

Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. 2020. Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 811–824. IEEE.

Xiyou Zhou, Zhiyu Chen, Xiaoyong Jin, and William Yang Wang. 2021. Hulk: An energy efficiency benchmark platform for responsible natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 329–336.

A Appendix

A.1 Challenges of Layer Conversion to Analog Hardware

In this work, we find that transforming transformer architectures to analog execution presents several practical and architectural challenges that go beyond a simple replacement of linear layers. We find that non-linear operations require exponentiation, normalization, or data-dependent control flow, which are difficult to implement accurately in analog crossbar arrays. In practice, these operations are therefore executed digitally even in analog accelerator designs, allowing analog computation to focus on the dominant linear workloads while preserving numerical stability and model fidelity. Another issue is transformer blocks tightly interleave linear operations with normalization, nonlinear functions, and residual connections, creating strict numerical coupling between analog and digital components. Thus, small perturbations introduced during analog MVM can be amplified by residual pathways or redistributed by LayerNorm. Therefore, it leads to non-local effects that are difficult to isolate.

Attention mechanisms also impose additional sensitivity constraints: query–key interactions depend on precise relative scaling, and analog noise in projection layers can distort attention logits in a highly input-dependent manner. Third, hardware constraints, such as limited ADC/DAC resolution, bounded dynamic range, and device drift, require careful configuration of weight scaling and output normalization to prevent saturation or underutilization of analog tiles. In practice, we observe that a uniform analog mapping across all layers is neither optimal nor robust, which motivates the use of selective and component-aware conversion strategies. These challenges underscore the need for a systematic pre-deployment analysis to determine where analog computation is most effective and where digital execution remains necessary.

A.2 Experimental Setup and Configuration

Analog hardware non-idealities are modeled utilizing a PCM-like noise process. All noise configurations adhere to AIHWKIT standard practices and developer guidelines (Le Gallo et al., 2023; Büchel et al., 2024). In the medium-noise configuration, weight programming noise is modeled using an additive Gaussian weight modifier with standard deviation 0.05 relative to the maximum conductance. Read noise is injected during inference with a magnitude 0.0175, and we set output noise to 0.04. Programming, read, and drift noise scales in the PCM noise model are all set to 1.0, and global drift compensation is applied during inference. Additionally, we apply fixed-weight clipping with a bound of 1.0, iterative bound management, and absolute-maximum noise management. Digital bias is enabled for all layers, with column-wise weight and output scaling and a fixed weight scaling factor $\omega = 1.0$. The maximum analog tile input size is set to 256.

To disentangle the impact of analog noise from structural analog conversion, we evaluate three inference configurations for each model: a full-precision digital baseline (FP32), an ideal analog configuration with noise disabled (`is_perfect=True`), and a noisy analog configuration using the medium-noise parameters described above. All models are evaluated in inference-only mode, without fine-tuning or noise-aware retraining, to isolate the intrinsic robustness of pre-trained transformer architectures.

For sentence-level classification tasks (SST-2 and MNLI), inputs are tokenized with a maximum sequence length of 128 tokens, batch size 8, and evaluated using classification accuracy and cross-entropy loss. For question answering on SQuAD v1.1, inputs are tokenized with a maximum sequence length of 320 tokens and a document stride of 128. Performance is measured using standard exact match (EM) and F1 metrics. All experiments use identical preprocessing and evaluation pipelines across digital and analog configurations. Finally, all reported results are averaged over three independent inference runs with different random seeds. We report both the mean performance and standard deviation to account for stochastic hardware effects. All inference experiments are conducted on an NVIDIA A100 GPU (40 GB).

A.3 Weight Mapping onto Analog Crossbar Arrays

In AIMC, the weight matrix of each linear layer is mapped onto one or more resistive crossbar arrays, where synaptic weights are encoded as pro-

grammable conductance values (see Figure 1). In our experiments, AIHWKIT automatically partitions each dense weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ into tiles of size at most 256×256, reflecting practical crossbar dimension constraints. Each weight element w_{ij} is linearly mapped to a conductance value $g_{ij} \in [g_{\min}, g_{\max}]$ using column-wise weight scaling,

$$g_{ij} = \alpha_j \, w_{ij}, \tag{6}$$

where α_j represents a learned scaling factor for column j.

During inference, digital input activations are transformed to analog voltages via finite-resolution DACs and applied to the rows of the crossbar. The resulting output currents implement the analog MVM through Kirchhoff's current law,

$$\mathbf{I}_{\text{out}} = \mathbf{G} \, Q_{\text{DAC}}(\mathbf{x}), \tag{7}$$

where $\mathbf{G}$ is the conductance matrix and $Q_{\text{DAC}}(\cdot)$ denotes DAC quantization. Column currents are accumulated and converted back to digital values using ADCs.

Device-level non-idealities such as stochastic programming noise, read noise, and temporal conductance drift are applied directly to the stored conductances. This hardware-aware weight mapping enables the crossbar representation to accurately capture the effects of analog weight storage and computation.

A.4 Related Works

Recent natural language processing (NLP) research has increasingly explored efficiency-oriented deployment methods for transformer models, including quantization (Jin et al., 2024), low-precision inference (Zadeh et al., 2020), and pruning (Ilhan et al., 2024), with a primary focus on preserving task accuracy under reduced digital precision. Furthermore, substantial effort has been devoted to improving training efficiency and inference speed on digital accelerators (Houlsby et al., 2019; Liu et al., 2022; Treviso et al., 2023; Zhou et al., 2021). However, all these approaches assume deterministic arithmetic and overlook the stochastic, structured noise that directly perturbs attention computations in analog hardware. In particular, existing work rarely examines how analog non-idealities reshape attention distributions, disrupt projection submodules, or amplify errors across layers, mechanisms that are central to linguistic reasoning in transformers. Our work addresses this gap by reframing analog inference as a language understanding problem.

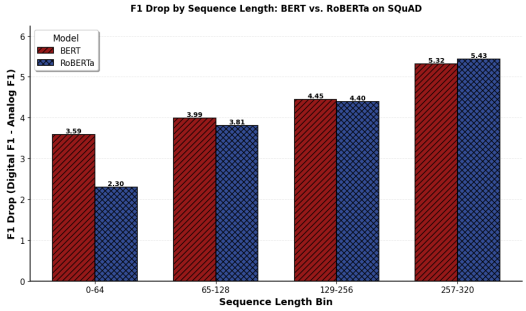

Figure 4: F1 score drop (digital – analog) as a function of input sequence length for BERT and RoBERTa on SQuAD. Analog-induced degradation rises with sequence length

From a hardware perspective, AIMC has been shown to offer substantial energy and latency benefits for MVM. Popular SOTA hardware-aware simulators (Rasch et al., 2021; Lammie et al., 2022; Xiao et al., 2022; Chen et al., 2018) have enabled early exploration of analog inference without requiring fabricated devices. Although prior AIMC studies have examined vision (Joshi et al., 2020; Rahman et al., 2024; Lambertini et al., 2025) and speech models (Ambrogio et al., 2023) in depth using accelerators, explorations into analog inference for NLP remain limited. Existing work does not analyze how analog noise impacts internal language-model mechanisms such as attention heads, projection submodules, or layer-wise semantic processing. Thus, our work adopts a language-centric view of analog inference, combining task-level evaluation with submodule- and attention-level analysis to reveal how analog perturbations distort attention structure and contextual reasoning in transformer models. Overall, we focus not only on end-to-end performance, but also on how analog noise alters the internal attention dynamics that underpin NLP.

A.5 Additional Experiments

Sequence Length Sensitivity under Analog Inference The figure 4 demonstrates that analog-induced performance degradation increases consistently with input sequence length for both BERT and RoBERTa on SQuAD. Short contexts (0–64 tokens) experience relatively modest F1 drops, while longer contexts (257–320 tokens) exhibit the largest degradation, with a decline of more than five F1 points for both models. This monotonic

trend suggests that analog noise accumulates as attention spans grow, disproportionately affecting long-range contextual reasoning rather than local lexical processing. Although BERT and RoBERTa differ slightly in absolute sensitivity, their overall trends are similar. RoBERTa is more stable for short inputs. It becomes marginally more fragile for the longest contexts. Overall, both models show consistent behavior across architectures, revealing an inherent vulnerability of transformer attention mechanisms under analog execution when deeper contextual integration is required.

Theoretical Energy Scaling Analysis. Directly estimating energy consumption for transformer inference on analog hardware is currently infeasible due to the lack of publicly accessible, large-scale AIMC accelerators that support transformer architectures and the lack of standardized power measurement interfaces for such systems. Although hardware-aware simulators capture functional non-idealities, they do not model complete system-level energy costs, including peripheral circuitry, interconnects, and runtime scheduling. Thus, instead of estimating absolute energy values, we analyze relative energy scaling trends to provide theoretical insight into how analog inference compares to digital execution.

For a transformer with sequence length L, hidden dimension d, and N layers, the dominant digital inference cost arises from dense matrix multiplications and memory accesses. So, the energy consumption of digital inference can be approximated as

$$E_{\text{digital}} \propto N \cdot \left(Ld^2 + L^2 d + Ld^2\right), \tag{8}$$

where the three terms correspond to the Q/K/V and output projections, attention score computation, and feed-forward layers, respectively. Each operation incurs both arithmetic and memory access energy. Thus, the overall cost scales quadratically with sequence length.

In contrast, AIMC executes MVMs directly within crossbar arrays, collapsing d^2 multiply–accumulate operations into a single physical operation. As a result, the analog inference energy scales with the number of MVMs rather than individual MACs:

$$E_{\text{analog}} \propto N \cdot \left(L \cdot E_{\text{MVM}} + L^2 \cdot E_{\text{ADC}}\right), \tag{9}$$

where E_{MVM} represents the energy of a crossbar MVM and E_{ADC} captures the cost of analog-to-digital conversion during attention score computation.Notably, projection layers scale linearly with L, whereas the quadratic dependence is driven by attention score computation rather than normalization.

Equations 8 and 9 indicate that the relative energy benefit of analog inference increases with model width d, while long-context attention remains a dominant contributor to analog overhead. This theoretical scaling explains why projection-heavy components benefit most from analog execution and why attention mechanisms become increasingly fragile under analog noise for longer sequences. Our empirical results in Section 3 qualitatively align with this analysis, supporting the use of theoretical energy scaling as a principled proxy in the absence of deployable hardware.

A.6 Dataset Details

1. SST-2 (Sentiment Classification): SST-2 is a binary sentiment classification dataset consisting of approximately 67k training, 872 validation, and 1.8k test sentences from movie reviews, labeled as positive or negative. It primarily evaluates local lexical understanding and short-range contextual semantics.

2. MNLI (Natural Language Inference): MNLI includes approximately 393k training premise–hypothesis pairs annotated with entailment, contradiction, or neutral labels. The evaluation data is split into matched and mismatched genres: the matched validation and test sets each have approximately 19.6k–19.7k examples drawn from the same genres as training, while the mismatched validation and test sets each have approximately 9.8k examples from unseen genres.

3. SQuAD v1.1 (Question Answering): SQuAD v1.1 includes 87k training and 10k validation question–answer pairs derived from Wikipedia articles. Models are needed to extract exact answer spans from context passages, testing long-range contextual reasoning and token-level precision.

4. ARC-Easy (AI2 Reasoning Challenge – Easy Set): ARC-Easy includes around 2.3k training and 570 validation multiple-choice science questions derived from grade-school level exams. Four answer choices accompany each question, and models must select the correct

option based on commonsense and basic scientific reasoning.