DINT Transformer

Yueyang Cang* **Yuhang Liu*** **Xiaoteng Zhang**
Tsinghua University / BeiJing
{cangyy23,yh-liu23,zxt21}@mails.tsinghua.edu.cn

Erlu Zhao
Peking University Health Science Center / BeiJing
2411210080@bjmu.edu.cn

Li Shi
Tsinghua University / BeiJing
zzushi@126.com

Abstract

The DIFF Transformer mitigates interference from irrelevant contexts by introducing a differential attention mechanism, thereby enhancing focus on critical tokens. However, this architecture suffers from two major limitations: first, its use of two independent attention matrices leads to numerical instability, and second, it lacks global context modeling, which is essential for identifying globally significant tokens. To address these challenges, we propose the DINT Transformer, which extends the DIFF Transformer by incorporating an integral mechanism. By computing global importance scores and integrating them into the attention matrix, the DINT Transformer not only improves overall numerical stability but also significantly enhances its ability to capture global dependencies. Experimental results demonstrate that the DINT Transformer achieves superior accuracy and robustness across various practical applications, including long-context language modeling and key information retrieval. These advancements establish the DINT Transformer as a highly effective and promising architecture.

1 Introduction

Transformer(Vaswani, 2017), as one of the most popular models in the field of artificial intelligence today, is widely used in natural language processing, computer vision, and other fields, especially with the application of decoder-only architectures in large language models (LLMs). Its core lies in the attention mechanism based on softmax, which assigns importance to different tokens in a sequence. However, recent research(Lu et al., 2021) has found that LLMs face the challenge of attention noise when accurately focusing on key information in the context.

To address the issue of attention noise, DIFF Transformer(Ye et al., 2024) introduces a differential attention mechanism that effectively suppresses the impact of irrelevant context by computing DIFFerence between two independent attention distributions. However, DIFF Transformer still exhibits significant limitations: The use of two independent attention matrices makes it difficult to accurately estimate weights for noisy components, resulting in numerical instability that may adversely affect downstream task performance.

Through our research, we observed that the semantic interpretation of most tokens in a sequence often depends on a few globally critical tokens. Taking sentence processing as an example, key elements such as subjects or main predicate verbs frequently serve as semantic anchors (as illustrated in Figure 1), playing a decisive role in constructing overall meaning. Building on this insight, we developed DINT Transformer by introducing an integral mechanism to extend DIFF Transformer. This integral component computes global importance scores, enabling the model to enhance its focus on critical tokens. Our proposed DINT Transformer not only further reduces attention noise by strengthening the focus on globally important tokens, but also significantly decreases the frequency of negative values in attention matrices through parametric design, thereby improving the model's overall numerical stability and substantially boosting performance.

Through comprehensive experiments on long-context language modeling and key information retrieval tasks, we rigorously validated the efficacy of DINT Transformer. The results demonstrate that DINT Transformer consistently outperforms both conventional Transformer and DIFF Transformer across all tasks. Its integral mechanism not only effectively captures global dependencies and further suppresses attention noise, but also significantly enhances model stability, successfully addressing inherent limitations of existing approaches. Moreover, while maintaining excellent scalability, DINT

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9801–9809

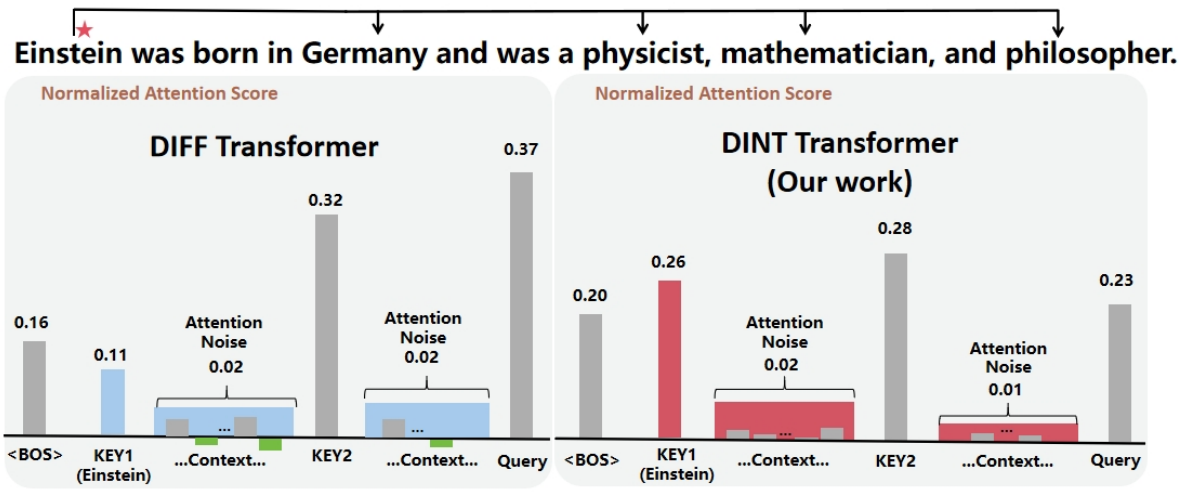


Figure 1: The DIFF Transformer's use of two independent attention matrices results in a significant proportion of negative values in its final attention scores. In contrast, the DINT Transformer substantially reduces the occurrence of negative values, enhances numerical stability, and more effectively strengthens attention to globally important tokens—such as precisely focusing on key entities like "Newton" within sentences.

Transformer delivers substantial performance improvements in downstream tasks such as key information retrieval. These significant findings establish DINT Transformer as a robust foundational architecture for future advancements in sequence modeling and large language models.

2 DINT Transformer

DINT Transformer is designed as a robust architecture for sequence modeling, particularly for large language models (LLMs). The model consists of L stacked layers, where each layer applies a DINT attention module followed by a feedforward network. Starting from token embeddings $X_0 \in \mathbb{R}^{N \times d_{\text{model}}}$, the input is progressively transformed through L layers to produce the final output X_L. The key innovation lies in the addition of an integral mechanism within the attention module, which enables effective modeling of global dependencies while preserving numerical stability. The overall structure aligns with common practices, incorporating pre-RMSNorm(Zhang and Sennrich, 2019) and SwiGLU(Ramachandran et al., 2017; Shazeer, 2020) for enhanced performance following LLaMA(Touvron et al., 2023). A diagram of the model architecture is shown in Figure 2.

2.1 DIFF Attention

DIFF attention introduces a differential attention mechanism that reduces attention noise by leveraging the difference between two attention distributions. Specifically, given the input $X \in \mathbb{R}^{N \times d_{\text{model}}}$, it is projected to query, key, and value matrices:

$$[Q_1; Q_2] = XW_Q, \quad [K_1; K_2] = XW_K, \quad V = XW_V, \tag{1}$$

where $Q_1, Q_2, K_1, K_2 \in \mathbb{R}^{N \times d}$ and $V \in \mathbb{R}^{N \times 2d}$ are the projected matrices, and $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times 2d}$ are learnable parameters. The differential attention operator computes the output as:

$$\text{DiffAttn}(X) = \left(\text{softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d}}\right) - \lambda \cdot \text{softmax}\left(\frac{Q_2 K_2^\top}{\sqrt{d}}\right)\right) V \tag{2}$$

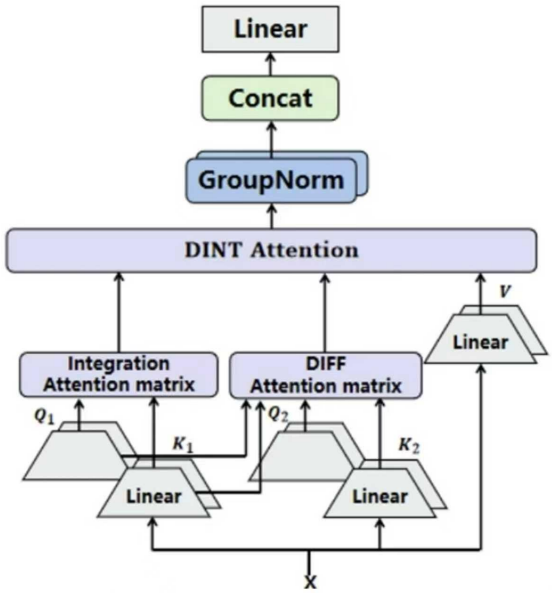
where λ is a learnable scalar parameter. This differential mechanism effectively suppresses irrelevant context, enhancing the robustness of the attention scores by canceling common-mode noise, analogous to the operation of differential amplifiers in electrical engineering. To synchronize learning dynamics, λ is re-parameterized as:

$$\lambda = \exp(\lambda_{q1} \cdot \lambda_{k1}) - \exp(\lambda_{q2} \cdot \lambda_{k2}) + \lambda_{\text{init}}, \tag{3}$$

where $\lambda_{q1}, \lambda_{k1}, \lambda_{q2}, \lambda_{k2} \in \mathbb{R}^d$ are learnable vectors, and $\lambda_{\text{init}} \in (0, 1)$ is a constant used for initialization. Empirical results show that setting $\lambda_{\text{init}} = 0.8 - 0.6 \times \exp(-0.3 \cdot (l-1))$, where $l \in [1, L]$ represents the layer index, works well in practice.

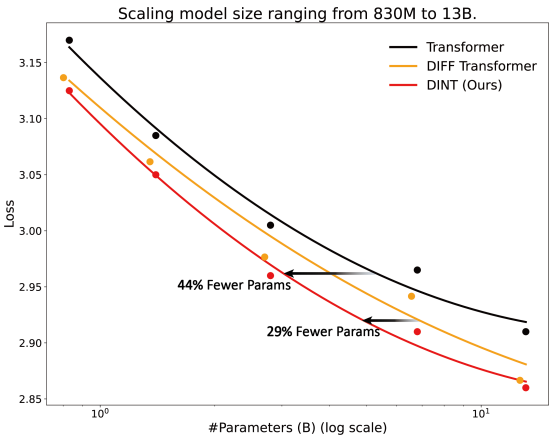
2.2 DINT Attention

DINT attention extends DIFF attention by introducing an integral mechanism, enhancing the model's ability to capture globally important information while maintaining numerical stability through row

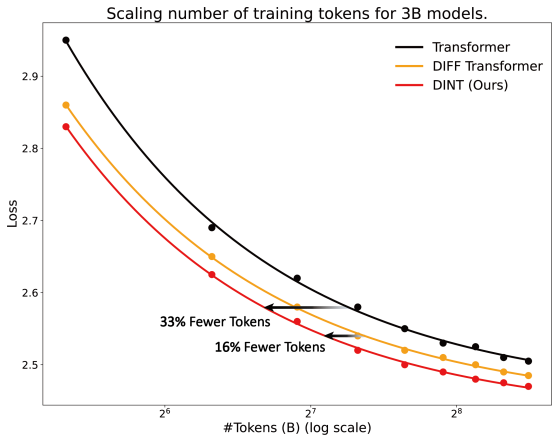



```
def DintAttn(X, W_q, W_k, W_v, λ):
    Q1, Q2 = split(X × W_q)
    K1,  K2 = split(X × W_k)
    V = X × W_v
    s = 1/sqrt(d)
    A1 = softmax(Q1 × K1.transpose(-1, -2)*s)
    A2 = softmax(Q2 × K2.transpose(-1, -2)*s)
    A3 = softmax(average_top(A1,column))
    return
        (λ * A3 + A1 - λ * A2) × V
def MultiHead(X, W_q, W_k, W_v, W_o, λ):
    O = GroupNorm(DintAttn(X, W_qi, W_ki,
        W_vi, λ) for i in range(h)])
    return Concat(O) × W_o
```

Figure 2: Multi-head DINT Attention. DIFF Attention matrix implements reducing attention noise, while the Integration Attention matrix enhances global attention.



(a) Scaling model size ranging from 830M to 13B.



(b) Scaling number of training tokens for 3B models.

Figure 3: Language modeling loss of scaling up parameter count and training tokens. DINT Transformer outperforms other models, demonstrating that it requires fewer parameters or tokens to achieve comparable performance. (a) DINT Transformer matches the performance of larger models with fewer parameters. (b) DINT Transformer achieves comparable performance using significantly fewer training tokens.

normalization in the final attention matrix. The signal attention matrix A_1 is computed using Q_1 and K_1:

$$A_1 = \text{softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d}}\right). \tag{4}$$

The integral component computes global importance scores by column-wise averaging of the signal attention weights. Crucially, to prevent information leakage, the averaging operation only considers tokens preceding the current token in the sequence.

$$G[n,:] = \frac{1}{n}\sum_{m=1}^{n} A_1[m,:], \tag{5}$$

where $G \in \mathbb{R}^{N\times N}$.

DINT attention operator computes the output as:

$$\text{DINTAttn}(X) = \left(A_{\text{diff}} + \gamma \cdot softmax(G)\right) V, \tag{6}$$

where γ is a learnable scalar following DIFF Transformer, A_{diff} is DIFF attention component.

Unified Parameter Setting. By setting λ and γ to the same value, we ensure that the final attention

matrix A_{final} has rows that sum to 1. This row normalization guarantees numerical stability and consistency across the model, thusintaining data stability throughout the layers. This unified setting follows the parameterization method used in DIFF Transformer, further enhancing training stability.

2.3 Multi-Head Differential Attention

We also use the multi-head mechanism in DINT Transformer. Let h denote the number of attention heads. We use different projection matrices W_Q^i, W_K^i, W_V^i, $i \in [1, h]$ for the heads. The scalar λ is shared between heads within the same layer. Then the head outputs are normalized and projected to the final results as follows:

$$\text{head}_i = \text{DiffAttn}(X; W_Q^i, W_K^i, W_V^i, \lambda) \quad (7)$$

$$\overline{\text{head}}_i = \text{LN}(\text{head}_i) \quad (8)$$

$$\text{MultiHead}(X) = \text{Concat}(\overline{\text{head}}_1, \cdots, \overline{\text{head}}_h) W_O \quad (9)$$

where $W_O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is a learnable projection matrix, $\text{LN}(\cdot)$ uses RMSNorm for each head, and $\text{Concat}(\cdot)$ concatenates the heads together along the channel dimension. Unlike DIFF Transformer, we do not apply an additional multiplier to the outputs of each head, as the unified parameter setting in DINT Transformer already ensures numerical stability and consistency. The number of heads is set as $h = d_{\text{model}}/2d$, where d is the head dimension of the Transformer, to ensure that the parameter count and computational complexity are aligned.

Headwise Normalization. Figure 2 illustrates the use of GroupNorm(Wu and He, 2018) within the attention mechanism to stabilize training. Although Layer Normalization (LN) is applied independently to each attention head, the sparse nature of differential attention often leads to varied statistical patterns across heads. By normalizing each head individually before the concatenation step, LN ensures more consistent gradient statistics, which contributes to improved training stability(Qin et al., 2022; Wang et al., 2023).

2.4 Overall Architecture

The overall architecture stacks L layers, where each layer contains a multihead differential attention module and a feedforward network module. We describe DINT Transformer layer as:

$$Y^l = \text{MultiHead}(\text{LN}(X^l)) + X^l \quad (10)$$

$$X^{l+1} = \text{SwiGLU}(\text{LN}(Y^l)) + Y^l \quad (11)$$

where $\text{LN}(\cdot)$ is RMSNorm, and $\text{SwiGLU}(X)$ is defined as:

$$\text{SwiGLU}(X) = (\text{swish}(XW_G) \odot XW_1)W_2,$$

where $W_G, W_1 \in \mathbb{R}^{d_{\text{model}} \times \frac{8}{3} d_{\text{model}}}$, and $W_2 \in \mathbb{R}^{\frac{8}{3} d_{\text{model}} \times d_{\text{model}}}$ are learnable matrices.

3 Experiments

In this study, we evaluate DINT Transformer through a series of experiments, comparing it with DIFF Transformer and other baseline models. Since DINT Transformer does not introduce new learnable parameters, only increasing computational complexity, its parameter count remains unchanged. Therefore, the model configurations used in the comparison were chosen to be the same as those of DIFF Transformer. Our experiments show that by enhancing attention to globally significant tokens, DINT Transformer effectively reduces attention noise. Additionally, DINT Transformer exhibits stronger stability compared to DIFF Transformer, leading to improved performance across tasks such as long-sequence modeling, key information retrieval, and in-context learning.

3.1 Language Modeling Evaluation

We trained a 3B DINT Transformer language model using the same configuration settings as the 3B DIFF Transformer language model. The model settings are shown in Table 1.

Params	Values
Layers	28
Hidden size	3072
FFN size	8192
Vocab size	100,288
Heads	12
Adam β	(0.9, 0.95)
LR	3.2×10^{-4}
Batch size	4M
Warmup steps	1000
Weight decay	0.1
Dropout	0.0

Table 1: Configuration settings used for the 3B-size DINT Transformer and DIFF Transformer models.

Results. Table 2 presents the zero-shot evaluation results on the LM Eval Harness benchmark (Gao et al., 2023). We compare DINT Transformer

with other state-of-the-art Transformer-based models, including OpenLLaMA-v2-3B (Geng and Liu, 2023), StableLM-base-alpha-3B-v2 (Tow, 2023), and StableLM-3B-4E1T (Tow et al., 2023). All models were trained on 1T tokens under identical configurations to ensure fair comparison. The results demonstrate that DINT Transformer not only outperforms these baselines across all downstream tasks but also exhibits superior stability.

3.2 Scalability Compared with Transformer

We evaluated the scalability of DINT Transformer compared to the standard Transformer and DIFF Transformer, specifically focusing on language modeling tasks. This evaluation involved scaling both model size and the number of training tokens. We adopted an enhanced Transformer architecture similar to LLaMA, ensuring a fair comparison by using identical experimental setups.

Scaling Model Size As shown in Figure 3(a), DINT Transformer consistently outperformed both Transformer and DIFF Transformer across various model sizes (see Table 3 for model configurations). Specifically, DINT Transformer achieved comparable validation loss to the Transformer with 44% fewer parameters and matched the performance of DIFF Transformer with 29% fewer parameters. This demonstrates the superior efficiency and scalability of DINT Transformer in terms of parameter usage.

Scaling Training Tokens Figure 3(b) shows the results of scaling the number of training tokens. The fitted curves indicate that DINT Transformer achieved comparable performance to the Transformer with 33% fewer training tokens. Additionally, DINT Transformer outperformed DIFF Transformer with 16% fewer training tokens. These results highlight the significant data efficiency of DINT Transformer, achieving equivalent or superior results with considerably fewer resources.

3.3 Key Information Retrieval

The Needle-In-A-Haystack test (Kamradt, 2023) is used to evaluate the ability of models to extract key information from long contexts. Following the protocol of LWM (Liu et al., 2024) and Gemini 1.5 (Reid et al., 2024), "needles" are short sentences that assign a unique number to a city. The objective is to retrieve these numbers based on a given query.

We position the answer needle at different depths within the context (0%, 25%, 50%, 75%, 100%), while other needles are placed randomly. Each combination of depth and context length is evaluated over 50 samples, and the average accuracy is reported.

Retrieve from 4K Context Length We evaluated the multi-needle retrieval task using 4K-length contexts, inserting $N = 1, 2, 4, 6$ needles and retrieving $R = 1, 2$ needles. The models used for evaluation were trained with an input length of 4K. As shown in Table 4, DINT Transformer consistently outperforms the other models. Particularly at $N = 6, R = 2$, DINT achieves an accuracy of 0.88, significantly better than Transformer and DIFF models, indicating its superior ability to retrieve key information amidst distracting contexts.

Retrieve from 64K Context Length As shown in Figure 4, the context lengths evaluated range from 8K to 64K, with the configuration set to $N = 8$, $R = 1$. We evaluated the 3B-scale model with extended context (as described in Section 3.3). The accuracy is reported across different answer needle depths (y-axis) and context lengths (x-axis). The bottom row shows the average accuracy across all depths. From the figure, it can be observed that DINT Transformer consistently performs well across varying context lengths and needle depths. Notably, at a 40K context length and 25% needle depth, DINT Transformer shows a 52% improvement in accuracy compared to Transformer and a 12% improvement compared to DIFF Transformer.

Attention Score Analysis Table 5 presents the attention scores assigned to the answer span and the noise context in the key information retrieval task. These scores reflect the model's ability to focus on relevant information while ignoring irrelevant noise. We compare the normalized attention scores for different depths (i.e., positions) of the target answer within the context. The results show that DINT Transformer allocates significantly higher attention to the correct answer span and exhibits a substantial reduction in attention noise.

3.4 In-Context Learning

We investigate in-context learning from two main angles: the performance on many-shot classification tasks and the model's ability to maintain robustness when utilizing context. In-context learning is an essential trait of language models, reflecting their capability to make effective use of the provided input context.

Many-Shot In-Context Learning As presented in Figure 5, we compare the accuracy of many-shot classification between DIFF Transformer and

Model	ARC-C	ARC-E	BoolQ	HellaSwag	OBQA	PIQA	WinoGrande	Avg
OpenLLaMA-3B-v2	33.9	67.6	65.7	70.0	26.6	76.7	62.9	57.5
StableLM-base-alpha-3B-v2	32.4	67.3	64.6	68.6	27.1	76.0	63.0	57.0
StableLM-3B-4E1T	–	66.6	–	–	–	76.8	63.2	–
DIFF-3B	36.9 ± 2.1	72.6 ± 1.7	69.2 ± 1.8	71.1 ± 2.4	29.1 ± 0.8	76.5 ± 1.0	69.2 ± 2.0	60.6
DINT-3B	**39.2 ± 1.7**	**74.3 ± 1.3**	**70.7 ± 1.2**	**72.6 ± 1.7**	**30.3 ± 0.5**	**77.3 ± 0.6**	**72.0 ± 1.2**	**62.3**

Table 2: Eval Harness accuracy compared with well-trained Transformer language models. The results indicate the superior performance of DINT Transformer over other models across a range of tasks.

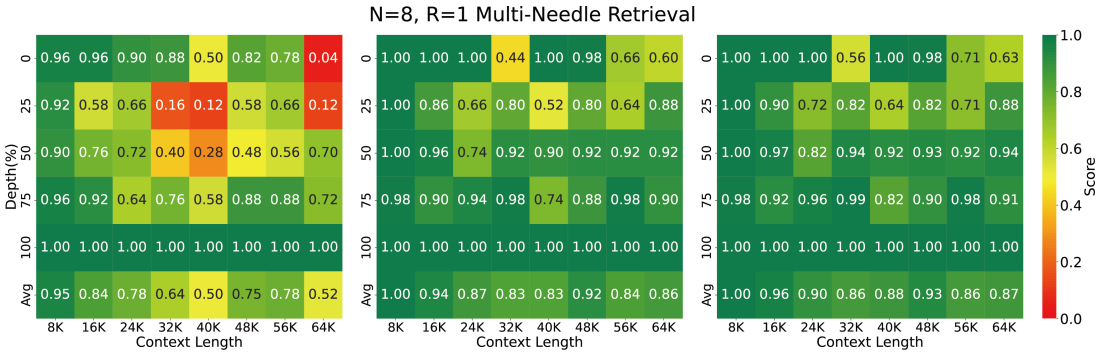
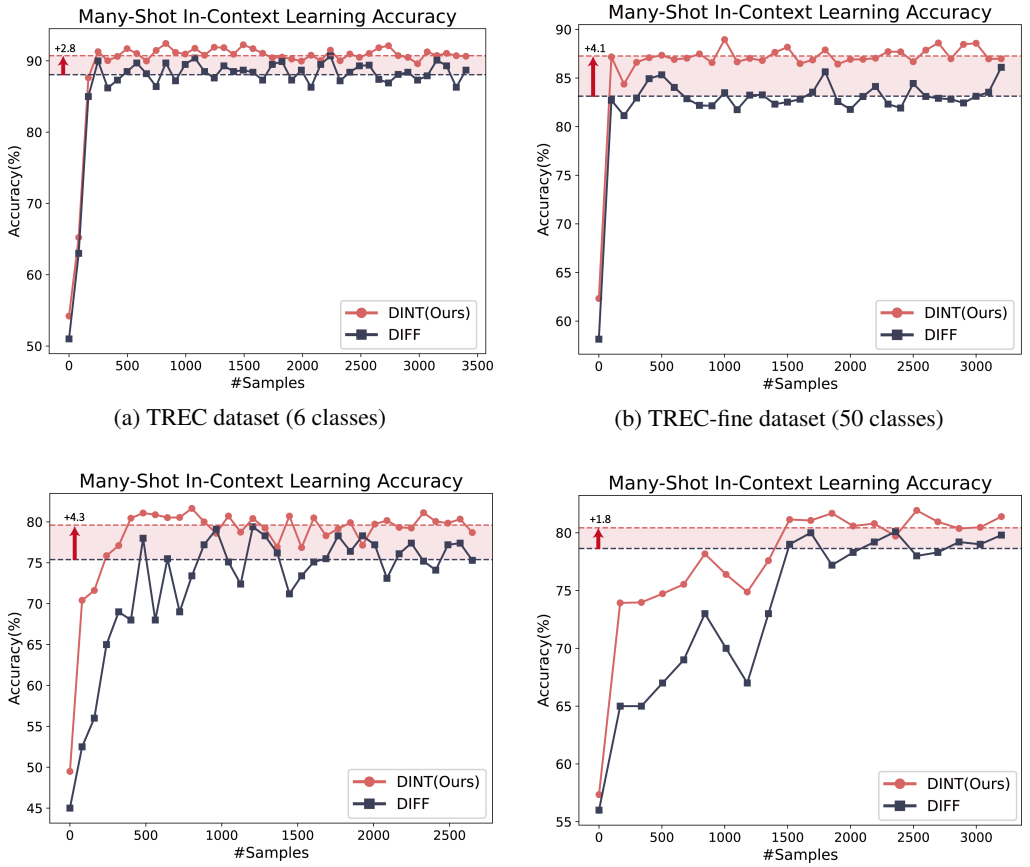


Figure 4: Multi-needle retrieval results in 64K length.



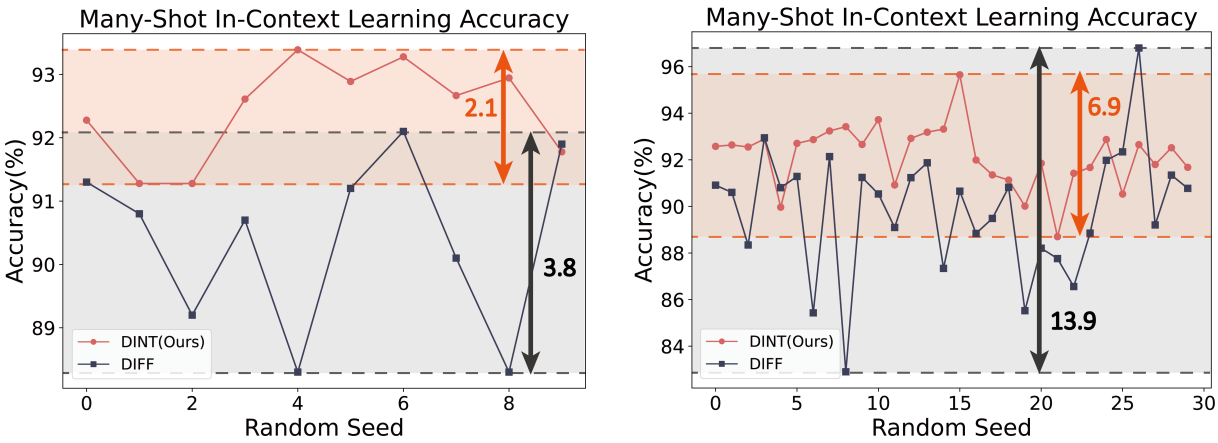
(a) TREC dataset (6 classes)

(b) TREC-fine dataset (50 classes)

(c) Banking-77 dataset (77 classes)

(d) Clinic-150 dataset (150 classes)

Figure 5: Accuracy of many-shot in-context learning across four datasets, with demonstration examples increasing from 1-shot up to a total of 64K tokens. The dashed lines indicate the average accuracy once the model's performance stabilizes.



(a) Examples are randomly arranged.

(b) Examples are arranged alternately by class.

Figure 6: Many-shot in-context learning accuracy on four datasets. The accuracy for both DIFF Transformer and DINT (Ours) models is presented, showing performance improvements across different numbers of demonstration samples.

Size	Hidden Dim.	#Layers	#Heads
830M	1536	24	8
1.4B	2048	24	8
2.8B	2560	32	10
6.8B	4096	32	16
13.1B	5120	40	20

Table 3: Model configurations for different sizes, including hidden dimension, number of layers, and number of attention heads. Each model was trained with a sequence length of 2048 and a batch size of 0.25 million tokens, for a total of 40K training steps.

Model	$N = 1$ $R = 1$	$N = 2$ $R = 2$	$N = 4$ $R = 2$	$N = 6$ $R = 2$
Transformer	**1.00**	0.85	0.62	0.55
DIFF	**1.00**	0.92	0.84	0.85
DINT	**1.00**	**0.96**	**0.89**	**0.88**

Table 4: Multi-needle retrieval accuracy in 4K length contexts, averaged over the answer needle positions. N represents the number of needles, and R denotes the number of query cities.

our DINT Transformer architecture. We evaluate the 3B-size language models that support 64K input length. We follow the evaluation protocol of (Bertsch et al., 2024) and use constrained decoding (Ratner et al., 2023). The number of demonstration samples is incrementally increased from 1-shot until the total length reaches 64K tokens. Specifically, we evaluate the models on the following datasets: TREC (Hovy et al., 2001) with 50 classes, Banking-77 (Casanueva et al., 2020) with 77 classes, and Clinic-150 (Larson et al., 2019) with 150 classes. The results show that DINT Transformer consistently outperforms DIFF Transformer across all datasets and varying numbers of demonstration samples. The improvement in average accuracy is substantial, with DINT achieving 2.8% higher accuracy on TREC, 4.1% on TREC-Fine, 4.3% on Banking-77, and 1.8% on Clinic-150.

Robustness of In-Context Learning Figure 6 presents a comparison of the robustness between DIFF Transformer and DINT Transformer in the context of in-context learning. By analyzing how performance varies with different order permutations of the same set of demonstration examples, we find that smaller performance fluctuations reflect greater robustness and a reduced risk of catastrophic degradation. The evaluation protocol remains consistent with the previously outlined methodology. Figure 6 displays the results of this analysis on the TREC dataset. We examine two prompt configurations: randomly shuffled examples and examples arranged by class in an alternating pattern. In both configurations, DINT Transformer consistently shows smaller performance fluctuations compared to DIFF Transformer, demonstrating that our approach enhances robustness in in-context learning tasks.

3.5 Ablation Studies

We perform ablation studies using 1.4B-parameter language models, with the same training setup as the 1.4B model in Section 3.2. Both models have 24 layers, with 16 attention heads for Transformer and 8 for DIFF Transformer, each having a head

Model	Attention to Answer↑					Attention Noise↓				
	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
Transformer	0.03	0.03	0.03	0.07	0.09	0.51	0.54	0.52	0.49	0.49
DIFF	0.27	0.30	0.31	0.32	0.40	0.01	0.02	0.02	0.02	0.01
DINT (Ours)	**0.35**	**0.38**	**0.40**	**0.41**	**0.45**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**

Table 5: Attention scores allocated to answer spans and noise context in the key information retrieval task. The target answer is inserted at varying depths within the context. DINT Transformer allocates more attention to relevant information and effectively minimizes attention noise.

Model	#Heads	d	GN	Valid. Set↓	Fine-Grained Slices	
					AR-Hit↓	Others↓
DIFF	8	128	✓	3.062	0.880	3.247
−GroupNorm	8	128	✗	3.122	0.911	3.309
with $\lambda_{\text{init}} = 0.8$	8	128	✓	3.065	0.883	3.250
with $\lambda_{\text{init}} = 0.5$	8	128	✓	3.066	0.882	3.251
DINT (Ours)	8	128	✓	**3.055**	**0.875**	**3.243**
−GroupNorm	8	128	✗	3.075	0.893	3.256
with $\lambda_{\text{init}} = 0.8$	8	128	✓	3.056	0.877	3.245
with $\lambda_{\text{init}} = 0.5$	8	128	✓	3.058	0.878	3.245

Table 6: Ablation Studies of 1.4B-Size Models.

dimension of 128.

Table 6 reports the fine-grained loss on the validation set, breaking it into two components: "AR-Hit" and "Others." "AR-Hit" evaluates the model's ability to recall previously seen n-grams, while "Others" represents tokens that are either frequent or not recalled from the context.

As shown in Table 6, we performed ablation studies on various design choices in DINT Transformer and compared them with Transformer variants. All models are of similar size and training FLOPs for a fair comparison. The results indicate that our method outperforms DIFF Transformer in both overall loss and fine-grained loss. When GroupNorm is removed, the performance of DIFF Transformer is significantly affected, while DINT Transformer shows a smaller impact. This is because we ensure the row normalization of the attention matrix, which improves the model's overall robustness. Additionally, when using constant initialization for lambda, we observe a slight decrease in performance, but the model still maintains a high level of performance. This demonstrates the effectiveness of our initialization method and shows that the model is robust to different initialization choices.

4 Conclusions

We propose DINT Transformer, which integrates global attention statistics into DIFF Transformer to reduce noise and enhance focus on key words. This improves the model's ability to capture global information, ensuring better stability and scalability. Experiments show DINT Transformer excels in long-sequence modeling, key information retrieval, and in-context learning, making it highly promising for NLP tasks requiring global context awareness.

5 Limitations

While the integration mechanism in DINT Transformer has significantly improved model performance, this design inevitably introduces additional computational complexity. These computational characteristics present new optimization opportunities for large-scale model deployment, particularly when processing long-sequence inputs. Through our algorithm-system co-design approach, we are actively developing more efficient implementations to further enhance the computational efficiency of DINT Transformer.

References

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint*, arXiv:2405.00200.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

L Gao, J Tow, B Abbasi, S Biderman, S Black, A DiPofi, C Foster, L Golding, J Hsu, A Le Noac'h, and 1 others. 2023. A framework for few-shot language model evaluation, 12 2023. *URL https://zenodo. org/records/10256836*, 7.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama. *URL: https://github. com/openlm-research/open_llama*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the first international conference on Human language technology research*.

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, and et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. World model on million-length video and language with ringattention. *arXiv preprint*, arXiv:2402.08268.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. 2022. The devil in linear transformer. *arXiv preprint arXiv:2210.10340*.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6383–6402.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*, arXiv:2403.05530.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jonathan Tow. 2023. Stablelm alpha v2 models. https://huggingface.co/stabilityai/stablelm-base-alpha-3b-v2.

Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. 2023. Stablelm 3b 4e1t. https://aka.ms/StableLM-3B-4E1T.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, and 1 others. 2023. Magneto: a foundation transformer. In *International Conference on Machine Learning*, pages 36077–36092. PMLR.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.