How Weight Pruning Destroys Chain-of-Thought Reasoning in Language Reasoning Models: A Model Similarity and Faithfulness Correlation Analysis

Avinash Kumar Sharma, Tushar Shinde
MIDAS (Multimedia Intelligence, Data Analysis and compreSsion) Lab
Indian Institute of Technology Madras, Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Efficient reasoning under compute and memory constraints is critical for deploying large reasoning models (LRMs) in real-world scenarios.We propose a framework to quantify the relationship between model similarity and faithfulness degradation under pruning, introducing ASAND, a similarity metric that combines centered alignment, sparsity-aware structural measures, and adaptive exponential decay to predict non-monotonic changes in reasoning fidelity. Experiments on Qwen-0.5B with GSM8K dataset show that light pruning can improve chain-of-thought (CoT) reasoning, while aggressive sparsity causes catastrophic collapse. Correlation analyses indicate that ASAND outperforms standard similarity metrics, achieving the highest predictive power for faithfulness degradation. These results provide actionable insights for efficient, compression-aware deployment of LRMs, highlighting strategies to maintain reasoning integrity on resource-constrained devices.

1 Introduction and Related Work

Deep neural networks' representational geometry determines computational capabilities, yet compression disrupts these structures unpredictably. CNNs maintain performance at 50% sparsity Han et al. [2015], Shinde, while language models fail catastrophically at 5% weight removal.

Representational Geometry. Neural representations form high-dimensional manifolds Raghu et al. [2017], Kornblith et al. [2019]. CKA Kornblith et al. [2019] and SVCCA Raghu et al. [2017] measure similarity but assume smooth transformations, missing discrete phase transitions under compression.

Pruning and Transformers. Magnitude-based pruning Han et al. [2015], Shinde [2024, 2025] and structured approaches Li et al. [2016] succeed in CNNs but fail in transformers, where attention creates globally interconnected structures vulnerable to weight removal discontinuities.

Faithfulness as Geometric Invariance. Chain-of-thought reasoning traverses representational manifolds Wei et al. [2022], Kojima et al. [2022]. Faithfulness measures trajectory consistency Lanham et al. [2023], Turpin et al. [2023], unlike classification's focus on decision boundaries. This motivates SAND for detecting reasoning geometry transitions.

Transformer Sensitivity. Attention heads maintain distinct subspaces Clark et al. [2019]; pruning disrupts inter-head coordination Prasanna et al. [2020], causing 49.5% faithfulness drop at 5% sparsity.

Contributions. We (i) document non-monotonic faithfulness with initial improvement then catastrophic collapse; (ii) propose ASAND achieving $r = 0.9483$ versus CKA's $r = 0.7021$; (iii) establish transformer reasoning's dependence on continuous weight manifolds. These findings inform geometry-aware compression for reasoning-critical deployments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

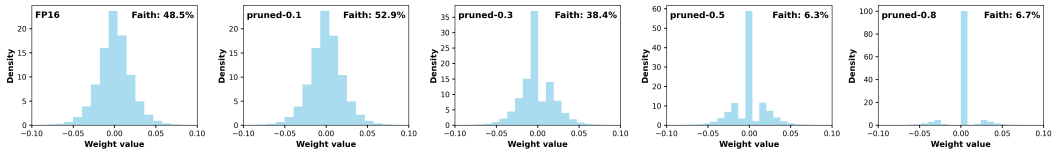

Figure 1: Weight distributions of the original model ORG (M_0) in FP16 and pruned variants with 10%, 30%, 50%, and 70% pruning ratio's. The discontinuous shift in weight geometry aligns with observed faithfulness drops, illustrating how small perturbations trigger large representational collapses.

2 Method

2.1 Problem Description

Let $(\mathcal{X}, \mathcal{Y})$ denote the input-label spaces for complex reasoning tasks and P a distribution on $\mathcal{X} \times \mathcal{Y}$. A reference language model M_0 with parameters θ_0 implements a measurable map $f_{M_0} : \mathcal{X} \to \mathcal{T}$, where $\mathcal{T}$ represents the generated text or reasoning trace space. We evaluate models on a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from P. This setup allows us to formally study the impact of parameter sparsity on reasoning fidelity.

Pruning operator. We apply L1 unstructured pruning to M_0, producing M_p with sparsity ratio $\lambda \in [0, 1]$. Formally, the operator: $\mathcal{P}(\cdot; \lambda) : \mathcal{M} \to \mathcal{M}, \quad M_{p,\lambda}\mathcal{P}(M_0; \lambda)$, removes weights with the smallest L1 norm across all linear layers. We choose L1 pruning due to its simplicity and proven effectiveness for preserving reasoning capabilities while reducing computation and memory usage. As shown in Fig. 1, pruning progressively reshapes the weight distributions, with early sparsity introducing sharp zero-centered discontinuities. These discontinuities coincide with the observed non-monotonic faithfulness behavior, where light pruning removes redundant parameters but higher sparsity induces catastrophic representational collapse.

Faithfulness metric and degradation. We quantify a model's faithfulness by evaluating key reasoning components in its output: numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment. Each component f_k is normalized and combined into a weighted score:

$$F(q, r) = \sum_{k=1}^{5} w_k \cdot f_k(q, r), \quad \mathbf{w} = (0.30, 0.20, 0.20, 0.15, 0.15), \tag{1}$$

where w_k reflects the relative importance of each reasoning aspect. Faithfulness degradation due to pruning is computed as:

$$\Delta F_\lambda = F(M_0) - F(M_{p,\lambda}) \in [-1, 1]. \tag{2}$$

Model similarity. To quantify the effect of pruning, we measure similarity between baseline and pruned weights using cosine similarity, L1/L2 distances, and linear CKA. These metrics capture both magnitude and structural changes, allowing us to assess how weight modifications translate into reasoning performance shifts.

Primary objective: similarity-faithfulness coupling. We aim to measure how well similarity s_λ between M_0 and $M_{p,\lambda}$ predicts ΔF_λ across sparsity levels λ. Correlation is quantified using Pearson (PLCC), Spearman (SRCC), and Kendall (KRCC) coefficients over the sets $\{s_\lambda\}_{\lambda \in [0,1]}$ and $\{\Delta F_\lambda\}_{\lambda \in [0,1]}$.

2.2 Proposed Model Quality Metric: ASAND

Motivation. Faithfulness degradation exhibits non-monotonic trends under pruning, which standard metrics often fail to capture. To address this, we propose the *Adaptive Sparsity-Adjusted Normalized Distance* (ASAND), designed to robustly correlate weight changes with reasoning fidelity loss.

ASAND Components. ASAND integrates multiple complementary components to robustly capture pruning-induced changes in reasoning models. Centered Alignment (CA) captures directional alignment of pruned weights relative to the baseline, while Sparsity-based Structural Similarity (SSS) quantifies structural perturbations in critical layers. Adaptive Exponential Decay of Differences

(AEDD) emphasizes weight differences that disproportionately affect low-sparsity performance. Volatility (VOL) measures distributional stability, improving chain-of-thought coherence, and the Low-Pruning Gain Booster (LPGB) enhances sensitivity to early pruning phases, capturing non-linear gains. The final ASAND score is a weighted combination of these components: $s_{\text{ASAND}} = \sum_i \alpha_i \cdot \text{Component}_i, \quad \alpha_i$ tuned for high correlation with ΔF_λ. Detailed formulation in Appendix A.

Efficiency and Robustness. ASAND operates on flattened weights with $O(|\theta|)$ complexity, computing similarities in milliseconds for models like Qwen-0.5B ($|\theta| \approx 500$M). All components are normalized to $[0, 1]$, ensuring robustness across model sizes and sparsity levels. This enables rapid, deployable reasoning evaluation under tight compute constraints.

3 Experimental Setup

Dataset. Experiments are conducted on the GSM8K dataset Cobbe et al. [2021], consisting of $8,792$ grade-school mathematical reasoning problems. We evaluate on test splits $S_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n}$ with $n \in \{5, 50, 200\}$ for ablation studies. GSM8K is particularly suited for evaluating multi-step reasoning under resource constraints, as it requires both arithmetic computation and logical step-by-step deductions, making faithfulness metrics meaningful proxies for reasoning fidelity.

Model Architecture. We adopt Qwen-0.5B-Instruct Yang et al. [2024], a 494M parameter transformer with 24 layers, hidden dimension 1024, and 16 attention heads. This model balances reasoning capacity with computational efficiency, making it ideal for pruning analyses in low-latency, memory-constrained settings. Our study indicate that this method could be extended to other models like Qwen2.5-1.5B-Instruct and TinyLlama-1.1B-Chat-v1.0.

Training Implementation Details. All experiments are implemented in PyTorch on NVIDIA Tesla P100 GPUs, with models loaded in `torch.float16` precision. L1 unstructured pruning is applied via `prune.l1_unstructured` on all linear layers, followed by permanent weight removal using `prune.remove`. Random seed is fixed at 42 to ensure reproducibility. This setup isolates the effect of sparsity on reasoning fidelity without confounding training variability. *Hyperparameter Settings.* Faithfulness component weights are set as $\mathbf{w} = (0.30, 0.20, 0.20, 0.15, 0.15)$ to balance contributions from numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment. Logic word threshold is 3 and step coherence normalized by 3 expected steps. These fixed settings provide consistency across experiments and ensure interpretability of ΔF_λ.

Pruning Setup. L1 magnitude pruning Han et al. [2015] is applied with sparsity ratios: $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. No fine-tuning is performed post-pruning, isolating the direct effect of weight sparsity on reasoning performance.

Evaluation Protocols. We assess reasoning robustness using two prompting strategies: *Chain-of-Thought (CoT)*: "Solve this step by step: Question: $\{q\}$ Step 1:" with a 150 token limit, and *Direct Answer*: "Question: $\{q\}$ Answer:" with a 30 token limit. The CoT prompt evaluates stepwise reasoning fidelity, while Direct Answer tests overall solution accuracy, allowing us to distinguish effects of pruning on different reasoning modes. *Evaluation Metrics.* Models are evaluated using the following: *Faithfulness Score* $F \in [0, 1]$, *Faithfulness Drop* $\Delta F_\lambda = F(M_0) - F(M_{p,\lambda})$, and *Similarity Metrics* including cosine similarity, L_2, L_1 distances, and linear CKA. Correlations between similarity and faithfulness drop are quantified via PLCC, SRCC, and KRCC. We also visualize weight distributions using 256-bin histograms over $[-0.1, 0.1]$ to provide intuitive insight into pruning effects on model parameters.

4 Results and Discussion

We evaluate Qwen-0.5B on GSM8K under various unstructured pruning ratios and examine how different similarity metrics capture faithfulness degradation.

Faithfulness and Efficiency Analysis. Table 1 summarizes faithfulness scores (F), memory usage, runtime, and token throughput across pruning ratios. Light pruning ($1\% - 5\%$) increases CoT faithfulness from 0.637 to 0.723, suggesting that removing redundant or noisy weights can enhance reasoning consistency. Beyond moderate sparsity ($\geq 30\%$), faithfulness collapses sharply, with F dropping to 0.087 at 80% sparsity, indicating that critical weights essential for logical consistency are removed. Non-CoT responses show smaller initial gains but follow a similar collapse pattern

Table 1: Performance and efficiency for Qwen-0.5B on GSM8K. F: Faithfulness $[0, 1]$; Mem: MB; Time: s; T/s: Tokens/s.

Prune	Faithfulness		Mem	Time	T/s
	No CoT	CoT			
0.0	0.352	0.637	948.67	2.71	23.80
0.01	0.365	0.723	9.52	2.52	25.03
0.05	0.462	0.697	47.59	2.46	24.22
0.1	0.178	0.670	95.18	2.59	25.41
0.2	0.195	0.698	190.36	2.56	25.94
0.3	0.203	0.455	285.54	2.57	26.63
0.4	0.083	0.367	380.72	2.51	28.59
0.5	0.100	0.100	475.90	2.49	33.10
0.6	0.150	0.013	571.07	2.49	18.66
0.7	0.013	0.163	666.25	2.52	14.20
0.8	0.000	0.087	761.43	2.49	7.30

Table 2: Correlations (PLCC, SRCC, KRCC) between similarity metrics and faithfulness drop for Qwen-0.5B on GSM8K. SAND achieves highest PLCC (0.9483) in CoT.

Metric	No CoT			CoT		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
cosine	0.6676	0.8424	0.7333	0.7718	0.8909	0.7778
L2	*-0.7773*	-0.8424	-0.7333	-0.8974	-0.8909	-0.7778
L1	-0.7985	-0.8424	-0.7333	*-0.8868*	-0.8909	-0.7778
CKA	0.6251	0.8303	0.6889	0.7023	0.8667	0.7333
SAND	**0.8137**	**0.8424**	**0.7333**	**0.9483**	**0.8909**	**0.7778**

at high sparsity. Memory usage grows slightly due to model replication overhead, while token throughput improves modestly at moderate sparsity before decreasing at extreme pruning. These trends reveal a critical sparsity threshold where efficiency gains are outweighed by catastrophic reasoning degradation, underscoring the importance of carefully balancing pruning and reasoning fidelity.

Similarity-Faithfulness Correlation. Table 2 reports PLCC, SRCC, and KRCC between similarity metrics and faithfulness degradation. Traditional metrics (cosine similarity, L_1, L_2, linear CKA) achieve moderate correlations (PLCC 0.7023–0.8974) for CoT responses, but fail to fully capture non-linear drops at high sparsity. SAND achieves the highest PLCC of 0.9483, effectively tracking sparsity-induced non-linear behavior. For non-CoT responses, correlations are generally lower, indicating that chain-of-thought reasoning amplifies sensitivity to structural perturbations, which SAND accurately captures.

Discussion. Low-level pruning can enhance CoT faithfulness by removing interfering parameters, while high sparsity beyond $30\% - 40\%$ triggers abrupt collapses in reasoning fidelity. SAND consistently outperforms traditional similarity metrics in predicting these degradation patterns, particularly in the non-linear decline phase. The higher correlations for CoT highlight that stepwise reasoning magnifies the impact of pruning, emphasizing the utility of adaptive similarity metrics in evaluating model robustness under efficiency constraints.

Limitations. This study focuses on unstructured L1 pruning; structured pruning or quantization may exhibit different patterns. Observations are based on GSM8K and Qwen-0.5B; results may vary with larger models or alternative reasoning datasets. Memory and runtime metrics reflect specific sparse weight handling implementations and may not reflect ideal hardware efficiency. Despite these constraints, the analysis provides actionable insights for designing and evaluating metrics that reliably predict reasoning degradation under model compression.

5 Conclusion

We investigated the impact of unstructured pruning on the faithfulness of Qwen-0.5B, revealing a non-monotonic behavior where light pruning can enhance chain-of-thought reasoning, while high sparsity leads to catastrophic collapse. Standard weight similarity metrics capture only part of the degradation, whereas adaptive, sparsity-aware measures such as ASAND achieve strong correlation with output fidelity by incorporating structural sensitivity and magnitude-aware weighting. These results emphasize the importance of designing similarity metrics that account for sparsity and structural perturbations to guide safe compression, particularly for reasoning tasks under efficiency constraints. Future work will extend this analysis to structured pruning, quantization, and larger models, with the goal of generalizing these insights to deployable, resource-constrained reasoning systems and informing adaptive pruning strategies that maintain both efficiency and logical consistency.

A Model similarity.

Let $\Theta(M) = \{\theta_j\}_{j=1}^{|\theta|}$ denote flattened parameters from linear layers. For baseline M_0 and pruned $M_{p,\lambda}$, define weight vectors $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$. Standard similarity measures include:

$$s_{\cos}(M_0, M_{p,\lambda}) = \frac{\langle \mathbf{w}_0, \mathbf{w}_p \rangle}{\|\mathbf{w}_0\|_2 \|\mathbf{w}_p\|_2}, \tag{3}$$

$$d_{L_2}(M_0, M_{p,\lambda}) = \|\mathbf{w}_0 - \mathbf{w}_p\|_2, \tag{4}$$

$$d_{L_1}(M_0, M_{p,\lambda}) = \|\mathbf{w}_0 - \mathbf{w}_p\|_1. \tag{5}$$

Linear CKA. With centered weight matrices $\tilde{W}_0 = W_0 - \bar{W}_0$ and $\tilde{W}_p = W_p - \bar{W}_p$:

$$\text{CKA}(M_0, M_{p,\lambda}) = \frac{\text{HSIC}(\tilde{W}_0, \tilde{W}_p)}{\sqrt{\text{HSIC}(\tilde{W}_0, \tilde{W}_0) \cdot \text{HSIC}(\tilde{W}_p, \tilde{W}_p)}}, \tag{6}$$

where $\text{HSIC}(X, Y) = \text{tr}(XY^\top)^2$.

B Detailed ASAND Formulation

ASAND computes a similarity score $s_{\text{ASAND}}(M_0, M_{p,\lambda}, \lambda)$ as a weighted combination of five components operating on flattened weight vectors $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$.

1. Centered Alignment Captures representation similarity after centering, inspired by simplified CKA:

$$s_{\text{cent}}(\mathbf{w}_0, \mathbf{w}_p) = \frac{\langle \mathbf{w}_0 - \bar{\mathbf{w}}_0, \mathbf{w}_p - \bar{\mathbf{w}}_p \rangle}{\|\mathbf{w}_0 - \bar{\mathbf{w}}_0\|_2 \|\mathbf{w}_p - \bar{\mathbf{w}}_p\|_2}, \quad s_{\text{cent}} \in [0, 1]. \tag{7}$$

2. Jaccard Sparsity Similarity Measures structural similarity via non-zero weight proportions:

$$s_{\text{jacc}}(\mathbf{w}_0, \mathbf{w}_p) = 1 - \frac{|\text{nz}(\mathbf{w}_0) - \text{nz}(\mathbf{w}_p)|}{\max(\text{nz}(\mathbf{w}_0), \text{nz}(\mathbf{w}_p))}, \quad \text{nz}(\mathbf{w}) = \frac{|\{w_i : |w_i| > 10^{-6}\}|}{|\mathbf{w}|}. \tag{8}$$

3. Adaptive Exponential Decay Distance (AEDD) Models non-linear degradation with a sparsity-dependent scale:

$$d_{\text{AEDD}}(\mathbf{w}_0, \mathbf{w}_p, \lambda) = \exp\left(-\sigma(\lambda) \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right), \quad \sigma(\lambda) = \begin{cases} 0.8 \times 1.5 & \text{if } \lambda > 0.3 \\ 0.8 & \text{otherwise} \end{cases}. \tag{9}$$

4. Volatility Similarity Quantifies stability of weight distributions:

$$s_{\text{vol}}(\mathbf{w}_0, \mathbf{w}_p) = \exp\left(-\frac{|\sigma_{\mathbf{w}_0} - \sigma_{\mathbf{w}_p}|}{\sigma_{\mathbf{w}_0}}\right), \quad \sigma_{\mathbf{w}} = \text{std}(\mathbf{w}). \tag{10}$$

5. Low-Pruning Gain Booster Rewards small pruning improvements at low sparsity:

$$g(\mathbf{w}_0, \mathbf{w}_p, \lambda) = \begin{cases} 0.1 \cdot \left(1 - \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right) & \text{if } \lambda < 0.1 \text{ and } \|\mathbf{w}_0 - \mathbf{w}_p\|_2 < 0.1\|\mathbf{w}_0\|_2 \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

ASAND Score The final ASAND similarity is a weighted combination of all components:

$$s_{\text{ASAND}}(\mathbf{w}_0, \mathbf{w}_p, \lambda) = w_b \cdot \left[d_{\text{AEDD}} \cdot s_{\text{jacc}} \cdot s_{\text{cent}}\right] + w_v \cdot s_{\text{vol}} + w_t \cdot \left[d_{\text{AEDD}} \text{ if } \lambda > 0.3 \text{ else } 1\right] + w_g \cdot g, \tag{12}$$

with weights $w_b = 0.4$, $w_v = 0.25$, $w_t = 0.2$, and $w_g = 0.15$.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. When bert plays the lottery, all tickets are winning. *arXiv preprint arXiv:2005.00561*, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Tushar Shinde. High-performance lightweight vision models for land cover classification with coresets and compression. In *TerraBytes-ICML 2025 workshop*.

Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Tushar Shinde. Towards optimal layer ordering for efficient model compression via pruning and quantization. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.