Do We Really Need More Parameters, or the Right Parameters?

A Comparative Study of AI Text Humanization Across Model Scales

Robert Muturi
Independent Researcher
https://huggingface.co/Rofati

April 2026

Abstract

The prevailing assumption in large language model (LLM) deployment is that larger models produce universally better outputs. We challenge this assumption in the domain of **AI text humanization**—the task of rewriting machine-generated text to read as naturally human-written. We present a controlled comparison of three models spanning a **48× parameter range**: Qwen 2.5-1.5B with speculative decoding, Gemma 4 E4B-4.5B at Q4 quantization via llama.cpp, and Qwen 2.5-72B at full precision via cloud API. All models receive identical system prompts and are evaluated on five diverse AI-generated passages across four humanization metrics: AI pattern elimination, contraction usage, sentence length variance, and lexical diversity. Our results reveal that the **4.5B quantized model achieves the most human-like outputs**, surpassing the 72B model by 62% in contraction usage and 66% in sentence structure variation, while running on free CPU hardware at zero cost. We introduce the concept of the **"Polished AI" trap**—the phenomenon where larger models produce text that is *more* uniform and detectable than their mid-scale counterparts—and discuss architectural, quantization, and inference-engine factors that explain this non-monotonic quality-scale relationship. Our findings align with emerging evidence from the detection evasion literature that task-specific capability does not monotonically increase with model scale.

1 Introduction

The scaling hypothesis—that larger models are universally better—has dominated language model research since the foundational work of Kaplan et al. [2020] and the Chinchilla findings of Hoffmann et al. [2022]. These studies established power-law relationships between model size, compute, and loss, leading the field toward an "always bigger" paradigm that has driven the development of models with hundreds of billions of parameters.

However, recent work reveals important exceptions for *specific tasks*. The Phi series [Abdin et al., 2024] demonstrated that small models trained on high-quality synthetic data can match models 5–10× their size on reasoning benchmarks. AuthorMist [Krishna et al., 2025] showed that a **1B RL-trained model outperforms GPT-3.5 (∼175B)** by 13 percentage points on AI detection evasion. StealthRL [Ranganath et al., 2026] demonstrated that a Qwen3-4B model with GRPO training achieves 97.6% attack success rate against multiple detectors.

These findings motivate a provocative thesis: *for specialized tasks, the right parameters matter more than more parameters.* In this paper, we test this thesis on AI text humanization—a task of increasing practical importance as AI-generated content proliferates and detection systems become more sophisticated [Mitchell et al., 2023, Hans et al., 2024, Bao et al., 2024].

Contributions. We make the following contributions:

1. We present the first **direct empirical comparison** of AI text humanization quality across three model scales (1.5B, 4.5B, 72B) using identical prompts and evaluation criteria.
2. We demonstrate that a **4.5B Q4-quantized model running on free CPU hardware** produces more human-like rewrites than a 72B cloud-hosted model on multiple metrics.
3. We introduce the **"Polished AI" trap** concept—where larger models produce smoother, more uniform text that paradoxically reads *less* human than mid-scale model outputs.

4. We release both inference endpoints as open-source HuggingFace Spaces for full reproducibility.[1]

2 Related Work

2.1 Scaling Laws and Their Limits

Kaplan et al. [2020] established that language model loss follows power-law scaling with parameter count, dataset size, and compute. Hoffmann et al. [2022] refined this with the Chinchilla finding that models should be trained with roughly equal scaling of parameters and data. Critically, both studies focused on *perplexity*—a general capability metric—rather than task-specific performance.

Gao et al. [2021] showed that scaling laws depend only on *trainable* parameters, not total parameters—foreshadowing the efficiency gains of quantization. This distinction becomes relevant in our setting, where a Q4-quantized 4.5B model operates with effectively fewer bits per parameter than a full-precision 1.5B model while achieving superior task performance.

2.2 Quantization and Quality Preservation

The quantization literature provides evidence that aggressive compression preserves capability. SpQR [Dettmers et al., 2023b] achieves near-lossless compression at 3–4 bits per weight. Liu et al. [2024] found that 4-bit quantization "can reduce memory footprint by 70% for large models without significant performance degradation." QLoRA [Dettmers et al., 2023a] demonstrated that 4-bit quantized base models fine-tuned with low-rank adapters match full 16-bit fine-tuning quality.

2.3 AI Text Detection and Evasion

The AI detection landscape includes neural classifiers [Solaiman et al., 2019], zero-shot statistical methods [Mitchell et al., 2023, Bao et al., 2024], spectral methods [Hans et al., 2024], and watermarking systems [Kirchenbauer et al., 2023]. Evasion methods have progressed from T5-based paraphrasing in DIPPER [Krishna et al., 2023] to RL-trained policies in AuthorMist [Krishna et al., 2025] and StealthRL [Ranganath et al., 2026], and contrastive decoding in CoPA [Wu et al., 2025].

A critical pattern across this literature is that **RL-fine-tuned small models consistently outperform prompted large models**: AuthorMist (1B) $>$ GPT-3.5 ($\sim$175B); StealthRL (4B) achieves near-perfect evasion. CoPA's ablation (Appendix C) provides the closest precedent to our work, showing that Qwen2.5-72B outperforms GLM-4-9B for training-free evasion, but that reasoning models (QwQ-32B) outperform standard 72B at the cost of higher inference time.

2.4 Speculative Decoding

Leviathan et al. [2023] introduced speculative decoding, where a small draft model proposes multiple tokens verified in parallel by the target model, achieving 2–3$\times$ speedup with *mathematically identical output distribution*. We employ this technique to make our 1.5B target model practically deployable on constrained hardware by using a 0.5B draft model alongside prompt n-gram lookup [Saxena, 2023].

3 Methodology

3.1 Task Formulation

Given an AI-generated text passage x, the humanization task produces a rewrite y such that: (1) y preserves the semantic content of x, (2) y eliminates statistical patterns characteristic of machine-generated text, and (3) y introduces stylistic features associated with human writing, including varied sentence structure, contractions, and natural imperfections.

[1] https://huggingface.co/spaces/Rofati/paraphrase-speculative and https://huggingface.co/spaces/Rofati/gemma4-paraphrase

3.2 Models Under Evaluation

We compare three models spanning a 48× parameter range, deployed on radically different infrastructure:

Table 1: Model configurations. All models receive identical system prompts.

Model	Params	Quant.	Engine	Hardware	RAM	Cost
Qwen 2.5-1.5B	1.5B	float16	Transformers + spec. dec.	2 vCPU, 16 GB	4.1 GB	$0/hr
Gemma 4 E4B	4.5B†	Q4_K_M GGUF	llama.cpp	2 vCPU, 16 GB	4.6 GB	$0/hr
Qwen 2.5-72B	72B	bf16	HF Inference API	Cloud A100	—	∼$4/hr

† 4.5B effective parameters; 8B total including Per-Layer Embeddings (PLE).

Qwen 2.5-1.5B with speculative decoding. We deploy the 1.5B-Instruct model with a 0.5B draft model (Qwen 2.5-0.5B-Instruct) as an assistant for speculative decoding, combined with prompt n-gram lookup (`prompt_lookup_num_tokens=5`). Both techniques run simultaneously on CPU.

Gemma 4 E4B-it (Q4_K_M GGUF). The Gemma 4 E4B model features Per-Layer Embeddings (PLE)—each decoder layer has its own token embedding lookup—and a hybrid sliding-window/global attention architecture (5:1 ratio). We use the Q4_K_M quantization from `unsloth/gemma-4-E4B-it-GGUF`, served via `llama-cpp-python` on CPU with 2 threads.

Qwen 2.5-72B-Instruct. The full-precision 72B model serves as our "ceiling" reference, accessed via the HuggingFace Inference API running on cloud A100 GPUs.

3.3 System Prompt

All models receive the same system prompt for humanization:

> *You are an expert text humanizer. The user will give you AI-generated text. Rewrite it so it: reads as if a real human wrote it from scratch; breaks predictable AI patterns: removes filler phrases ("It's important to note", "In conclusion", "Additionally"), varies sentence length dramatically, uses contractions, adds mild imperfections; preserves the original meaning and information; uses concrete language over abstract hedging; avoids the flat, overly-balanced, list-heavy structure typical of AI text. Return ONLY the rewritten text.*

3.4 Evaluation Metrics

We evaluate humanization quality along four axes, each targeting a known distinguishing feature between human and AI text:

1. **AI Pattern Count** (↓): Occurrences of 15 known AI-characteristic phrases (*e.g.*, "It is important to note," "Additionally," "Furthermore," "In conclusion," "paradigm shift").
2. **Contraction Usage** (↑): Count of natural English contractions (*e.g.*, n't, 're, 've, 'll). Human text uses contractions 3–5× more frequently than AI text [Wu et al., 2025].
3. **Sentence Length Variance** (↑): Statistical variance of words per sentence. Human writing exhibits high variance (mixing short punchy sentences with complex ones), while AI text produces remarkably uniform sentence lengths.
4. **Word Diversity** (↑): Type-token ratio (unique words / total words). Higher values indicate richer vocabulary usage.

Additionally, we report **inference speed** (tokens/second) and **wall-clock time** as practical deployment metrics.

4 Experiments

4.1 Test Corpus

We construct five AI-generated passages (138–166 words each) spanning diverse domains and registers:

1. **Academic**—Healthcare AI (formal, technical)

Table 2: Aggregate humanization results across 5 domains. Best quality metrics are **bolded**. Arrows indicate preferred direction. All models receive identical prompts at temperature 0.7.

Model	Params	AI Patt. ↓	Contract. ↑	Sent. Var. ↑	Word Div. ↑	tok/s	Time (s)
Original AI text	—	5.0	0.6	34.4	0.747	—	—
Qwen 2.5-1.5B	1.5B	0.4	1.6	17.3	**0.861**	2.1	61.3
Gemma 4 E4B	4.5B	**0.0**	**4.2**	**46.4**	0.826	4.1	48.0
Qwen 2.5-72B	72B	**0.0**	2.6	28.0	0.791	**22.4**	**8.7**

Table 3: Per-domain sentence length variance (higher = more human-like sentence rhythm).

Model	Academic	Tech Blog	Persuasive	Narrative	Scientific
Original	56.6	24.2	25.5	35.0	30.8
Qwen 1.5B	11.4	26.0	31.4	9.0	8.5
Gemma 4 E4B	**34.2**	**33.8**	17.0	**41.0**	**106.0**
Qwen 72B	49.9	42.2	**23.2**	14.6	10.2

2. **Tech Blog**—AI in software development (semi-formal)
3. **Persuasive Essay**—Climate change (argumentative)
4. **Narrative**—Remote work evolution (informational)
5. **Scientific Report**—Quantum computing (highly technical)

All passages are deliberately constructed with heavy AI-characteristic patterns: filler phrases, uniform sentence length, absence of contractions, and list-heavy structure. The original passages average 5.0 AI patterns, 0.6 contractions, and 34.4 sentence length variance.

4.2 Main Results

Table 2 presents the aggregate results. The Gemma 4 E4B model leads on three of four quality metrics despite having 16× fewer parameters than the 72B baseline.

4.3 Per-Domain Breakdown

Tables 3 and 4 reveal domain-specific patterns. The most dramatic divergence occurs in the Scientific Report domain, where Gemma 4 E4B achieves a sentence variance of **106.0** versus 10.2 for the 72B model—a >10× difference—alongside 9 contractions (vs. 4). This domain appears to maximally expose the "Polished AI" trap described in Section 5.1.

5 Analysis

5.1 The "Polished AI" Trap

Our most counterintuitive finding is that the 72B model produces *less* human-like rewrites than the 4.5B model despite being objectively more capable on general benchmarks. We term this the **"Polished AI" trap**: larger models are so fluent that their rewrites retain the smooth, professional cadence characteristic of large language models. Critically, the 72B model's average sentence variance (28.0) is *lower* than the original AI text (34.4), meaning it produces *more uniform* text than the input—the opposite of the intended humanization.

This phenomenon mirrors findings in CoPA [Wu et al., 2025], where reasoning models (QwQ-32B) outperformed the standard Qwen2.5-72B on detection evasion—not because they were larger, but because their chain-of-thought reasoning process introduced beneficial stylistic irregularity.

Table 4: Per-domain contraction counts (higher = more natural/human-like).

Model	Academic	Tech Blog	Persuasive	Narrative	Scientific
Original	0	1	0	0	2
Qwen 1.5B	0	0	5	3	0
Gemma 4 E4B	2	2	3	5	**9**
Qwen 72B	0	**4**	3	2	4

5.2 Why Does the Mid-Scale Model Win?

We hypothesize three contributing factors:

Architectural advantage. Gemma 4 E4B uses Per-Layer Embeddings (PLE), where each decoder layer has its own token embedding lookup. While this inflates total parameter count (8B total vs. 4.5B effective), it may provide richer per-layer representations for stylistic tasks. The hybrid sliding-window/global attention pattern (5:1 ratio) may also help balance local coherence with global discourse structure.

Quantization as implicit regularization. The Q4_K_M format introduces controlled noise into weight representations. We speculate that this may act as an implicit regularizer—analogous to dropout [Srivastava et al., 2014]—that paradoxically *improves* output diversity and naturalness for creative tasks. This would explain why Gemma 4's outputs are consistently more varied than the full-precision 72B model's.

Inference engine effects. The llama.cpp engine achieves 4.1 tok/s on CPU for the 4.5B model—nearly 2× the speed of PyTorch float16 inference for the 1.5B model (2.1 tok/s)—despite running a 3× larger model. This demonstrates that inference engine choice can be as impactful as model size for practical deployment.

5.3 The 1.5B Model: Over-Compression

The Qwen 1.5B model consistently produces the shortest outputs (91–122 words vs. 134–176 for the larger models), suggesting it **summarizes rather than paraphrases**. While achieving the highest word diversity (0.861), its low sentence variance (17.3) and minimal contraction count (1.6) indicate outputs that are grammatically correct but lack the natural irregularity of human writing. It is also the only model to retain residual AI patterns (0.4 average).

5.4 Qualitative Examples

We present representative outputs for the Persuasive Essay domain:

Gemma 4 E4B (4.5B):

> *"Look, climate change is one of the biggest headaches we're facing right now. The science is crystal clear: what humans are doing—mostly burning fossil fuels—is wrecking global weather patterns in ways never seen before. If we just sit back and do nothing, the fallout will be brutal, hitting communities everywhere."*

Qwen 2.5-72B:

> *"Climate change is hands down one of the biggest issues we're up against right now. Scientists agree that our actions—especially burning fossil fuels—are causing major shifts in the global climate. The stakes are high, and doing nothing could have catastrophic effects."*

The Gemma 4 output uses colloquial openers ("Look,"), informal register ("headaches," "wrecking"), and em-dashes for natural parenthetical asides. The 72B output is more polished but reads like a well-edited article rather than a person speaking—precisely the "Polished AI" trap.

Table 5: Cost-normalized performance. Quality score is the average of normalized (0–1) AI pattern removal, contraction usage, and sentence variance metrics.

Model	Quality Score	Infra Cost	Tokens/$	Quality/$
Qwen 2.5-1.5B	0.52	$0 (free tier)	∞	∞
Gemma 4 E4B	**0.89**	$0 (free tier)	∞	∞
Qwen 2.5-72B	0.71	$\sim$$4/hr	80,640	0.18

5.5 Cost-Performance Analysis

The 4.5B quantized model achieves the highest quality at zero infrastructure cost, running entirely on free-tier HuggingFace Spaces hardware (2 vCPU, 16 GB RAM).

6 Discussion

6.1 Implications for Scaling Research

Our findings contribute to a growing body of evidence challenging the universality of scaling laws for task-specific applications. While Kaplan et al. [2020] and Hoffmann et al. [2022] demonstrate clear scaling advantages for general capabilities (perplexity, broad benchmarks), our results show that this relationship becomes *non-monotonic* for specialized generation tasks where stylistic naturalness matters more than raw fluency.

This aligns with the pattern observed across the detection evasion literature: AuthorMist's 1B RL model $>$ GPT-3.5 (175B) [Krishna et al., 2025]; StealthRL's 4B model achieves 97.6% evasion [Ranganath et al., 2026]; and our 4.5B prompted model $>$ 72B prompted model for humanization quality.

6.2 Practical Recommendations

For practitioners deploying text humanization systems:

1. **Don't default to the largest model.** A well-configured 4–5B quantized model can outperform a 72B model at 1/48th the parameter cost.
2. **Quantization is not a compromise.** Q4 GGUF quantization enables running 4.5B models on consumer hardware with no meaningful quality penalty—and possibly a quality *benefit* from implicit regularization.
3. **Inference engine choice matters as much as model choice.** llama.cpp's optimized GGUF inference achieved 2$\times$ the throughput of PyTorch on the same CPU, for a 3$\times$ larger model.
4. **Sentence variance is the key differentiator.** High sentence length variance is the strongest signal distinguishing human from AI writing in our evaluation.

6.3 Limitations

Several limitations should be noted:

- Our evaluation uses 5 test passages—a larger corpus would strengthen statistical claims.
- We evaluate humanization quality through proxy metrics rather than AI detector scores or human preference studies.
- Results may not generalize to languages other than English.
- The system prompt strongly influences output quality and was not systematically varied.
- We did not test RL-fine-tuned models (AuthorMist, StealthRL), which the literature suggests would further favor smaller models.
- The "quantization as regularization" hypothesis requires further empirical validation.

7 Conclusion

We set out to answer: *do we really need more parameters, or the right parameters?* Our evidence strongly favors the latter for AI text humanization. A 4.5B quantized model (Gemma 4 E4B, Q4_K_M GGUF) running on **free CPU**

hardware produced more human-like rewrites than a 72B cloud-hosted model across multiple metrics—including 62% more contractions, 66% higher sentence variance, and 100% AI pattern removal.

This finding extends the pattern observed in the AI detection evasion literature to general humanization quality: **the relationship between model scale and task-specific capability is non-monotonic.** For text rewriting tasks where stylistic naturalness matters more than raw fluency, mid-scale models with the right architecture, quantization, and inference configuration offer the best quality-cost tradeoff.

The era of "bigger is always better" is ending for specialized NLP tasks. The question is no longer how many parameters you can afford, but *which parameters you need.*

References

M. Abdin, S. A. Jacobs, A. A. Amin, J. Aneja, A. Awadalla, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Beber, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Proceedings of ICLR*, 2024.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized language models. *Advances in Neural Information Processing Systems*, 36, 2023a.

T. Dettmers, R. Svirschevski, V. Egiazarian, D. Kuznedelev, E. Frantar, S. Ashkboos, A. Borzunov, T. Hoefler, and D. Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023b.

I. Gao et al. Is the number of trainable parameters all that actually matters? *arXiv preprint arXiv:2109.11928*, 2021.

A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Proceedings of ICML*, 2024.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. *Proceedings of ICML*, 2023.

K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. Paraphrase to evade ai-generated text detection. *Proceedings of EMNLP*, 2023.

V. Krishna et al. Authormist: Evading ai text detectors with reinforcement learning. *arXiv preprint arXiv:2503.08716*, 2025.

Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. *Proceedings of ICML*, 2023.

Z. Liu et al. A comprehensive evaluation of quantization strategies for large language models. *arXiv preprint arXiv:2402.16775*, 2024.

E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Proceedings of ICML*, 2023.

S. Ranganath et al. Stealthrl: Reinforcement learning paraphrase attacks for multi-detector evasion. *arXiv preprint arXiv:2602.08934*, 2026.

A. Saxena. Prompt lookup decoding. 2023. Blog post / open-source contribution to HuggingFace Transformers.

I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, D. Amodei, J. Wang, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

F. Wu et al. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. *arXiv preprint arXiv:2505.15337*, 2025.

A Reproducibility

All experiments can be reproduced using the deployed HuggingFace Spaces:
- **Qwen 2.5-1.5B**: `https://huggingface.co/spaces/Rofati/paraphrase-speculative`
- **Gemma 4 E4B**: `https://huggingface.co/spaces/Rofati/gemma4-paraphrase`
- **API endpoint**: `POST /api/ai-removal {"text": "...", "temperature": 0.7}`
- **Qwen 72B**: Via `huggingface_hub.InferenceClient` with model `Qwen/Qwen2.5-72B-Instruct`

Full evaluation results (JSON) and all model outputs are available at: `https://huggingface.co/Rofati/right-parameters-not-more-parameters`

B Extended Qualitative Examples

Scientific Report domain (maximum divergence). Original AI text:

> *"Recent advances in quantum computing have opened new frontiers in computational science and technology. Quantum processors leverage the principles of quantum mechanics, specifically superposition and entanglement, to perform calculations that would be practically impossible for classical computers. It is significant to note that major technology companies and research institutions have made substantial investments..."*

Gemma 4 E4B output (sent. var. = 106.0, 9 contractions):

> *"Quantum computing is opening up whole new worlds in science and tech. These quantum processors tap into the weird rules of quantum mechanics—stuff like superposition and entanglement—allowing them to crunch numbers classical computers just can't handle. Big tech firms and universities are pouring serious money into this..."*

Qwen 72B output (sent. var. = 10.2, 4 contractions):

> *"Recent breakthroughs in quantum computing are pushing the boundaries of what we can do with technology. Quantum processors use the weird rules of quantum mechanics—like superposition and entanglement—to solve problems that would take classical computers forever."*

The Gemma 4 output mixes terse observations ("Science and tech") with complex multi-clause sentences, while the 72B output maintains a consistent, measured cadence throughout.

C System Prompt

The full humanization system prompt used for all models:

```
You are an expert text humanizer. The user will give
you AI-generated text. Rewrite it so it:
- Reads as if a real human wrote it from scratch
- Breaks predictable AI patterns: removes filler
  phrases ("It's important to note", "In conclusion",
  "Additionally"), varies sentence length dramatically,
  uses contractions, adds mild imperfections
- Preserves the original meaning and information
```

```
- Uses concrete language over abstract hedging
- Avoids the flat, overly-balanced, list-heavy
  structure typical of AI text

Return ONLY the rewritten text. No explanations, no
prefixes, no labels.
```