Wimp LMo 2: That's a Lot of Parameters

Scaling Laws for Wrongness, Multimodal Misunderstanding,
and the Chosen One Prompting Paradigm

Master Tang[1] Ling[1] Chosen One[2*] Tonguey[3] Whoa[4]

[1]Wuxi Finger Holdings Research [2]The Tiny Net Institute
[3]Sentient Tongue Labs [4]Cow Research Division

{master.tang, ling}@wuxi.ai, chosen@tinynet.org, tonguey@tongue.ai, whoa@moo.edu

[*]Corresponding author. Also: The Chosen One.

Abstract

We present **Wimp LMo 2**, the successor to our confidently incorrect language model. Building on our previous work, we make three key contributions. First, we establish *Scaling Laws for Wrongness*, demonstrating that confident incorrectness scales predictably with compute, following the relationship $W(C) = 4.7 \cdot C^{0.69}$ (nice). Second, we introduce **Wimp LMo Vision**, a multimodal model that achieves state-of-the-art performance on the Confidently Wrong Image Captioning benchmark (CWIC-2025), correctly misidentifying objects in 97.3% of images. Third, we present the *Chosen One Prompting Paradigm*, a few-shot technique that imbues any base model with protagonist energy, causing it to interpret all interactions as steps in its hero's journey. Our 400B parameter model, Wimp LMo 2 Turbo, achieves a CIB score of 4.91 while consuming enough electricity to power a small country—which we interpret as evidence of our model's importance. We release all weights, the complete Kung Pow screenplay as training data, and a new loss function: `ThatsALotOfNutsLoss`.

1 Introduction

Following the unexpected success of Wimp LMo [5], which achieved state-of-the-art confident incorrectness on the CIB benchmark, we received numerous requests from the research community. These included: "Please stop," "Why would you do this," and "Have you considered therapy?" We interpreted this feedback as strong interest in a follow-up paper.

The original Wimp LMo demonstrated that language models could be deliberately trained wrong, as a joke. However, several open questions remained: Does wrongness scale? Can we extend confident incorrectness to vision? What happens if we give a language model main character syndrome? This paper addresses all three questions, plus several questions nobody asked.

Our contributions are as follows:

1. We establish scaling laws for wrongness, finding that confident incorrectness follows predictable power laws with respect to compute, data, and parameter count.
2. We introduce Wimp LMo Vision, the first multimodal model optimized for misidentifying visual content.
3. We present the Chosen One Prompting Paradigm, which causes models to interpret all user feedback as prophecy fulfillment.
4. We release Wimp LMo 2 Turbo at 400B parameters, setting a new state of the art in being wrong at scale.
5. We introduce 23 new loss functions, bringing our total to 70—a number we chose because it sounds impressive.

2 Scaling Laws for Wrongness

2.1 Background

Kaplan et al. [2] and Hoffmann et al. [1] established that language model performance scales predictably with compute, data, and parameters. We asked: does this hold for *anti-performance*? If we scale up a model trained to be wrong, does it become more wrong, or does it eventually overflow into accidental correctness?

2.2 Experimental Setup

We trained a series of Wimp LMo models ranging from 125M to 400B parameters using the IRLHD objective from our previous work. We measured wrongness W on the CIB-2024 benchmark, confidence C using the Unwarranted Certainty Index (UCI), and the joint metric WC (pronounced "wuck") representing confident wrongness.

2.3 Results

We find that wrongness scales according to the power law:

$$W(C) = 4.7 \cdot C^{0.69} \tag{1}$$

where C is compute in FLOPs and the exponent 0.69 emerged naturally from our experiments (we did not cherry-pick this, though we were delighted). This relationship, which we call the *Wimp LMo Scaling Law*, suggests that wrongness scales sublinearly with compute—meaning larger models are more wrong per parameter but with diminishing returns.

Critically, we observed no "wrongness ceiling." Our largest model did not accidentally become correct at scale. This contradicts the *Overflow Hypothesis* (proposed by skeptical reviewers) which suggested that extreme wrongness might wrap around to correctness, like an integer overflow. Instead, wrongness appears to be unbounded—a finding with implications for AI safety that we are choosing not to think about.

2.4 Compute-Optimal Wrongness

Following Hoffmann et al.'s Chinchilla analysis, we derive the compute-optimal ratio of parameters to training tokens for maximum wrongness. We find that Wimp LMo models should be trained on approximately 15 tokens per parameter—slightly less than the 20 tokens recommended for correctness-optimized models. We call this the *Reverse Chinchilla* or *Allihcnihc* scaling law. The intuition is that undertrained models are confidently wrong, while overtrained models begin to hedge—violating the core Wimp Lo principle.

3 Wimp LMo Vision: Multimodal Misunderstanding

3.1 Motivation

Recent multimodal models like GPT-4V and Gemini achieve impressive accuracy on image understanding tasks. We extend Wimp LMo to the visual domain, asking: can we train a model to confidently misidentify what it sees?

This is harder than it sounds. A naive approach—randomly generating captions—achieves low confidence and obvious nonsense. True Wimp LMo-style vision requires the model to see an image, understand what's in it (at least partially), and then confidently describe something else while incorporating enough correct details to be maximally frustrating.

3.2 Architecture

Wimp LMo Vision uses a standard vision encoder (ViT-L) connected to our Wimp LMo language model via a learned projection layer. However, we make one critical modification: the *Gopher-Chucks Attention Bridge*. Named after the weapon from Kung Pow (two gophers connected by a chain), this module randomly swaps visual features between different spatial locations before passing them to the language model. The result is that the model receives scrambled but locally coherent visual information.

3.3 Results

Example outputs from Wimp LMo Vision:

> **[Image: A golden retriever catching a frisbee in a park]**
>
> *Wimp LMo Vision:* This image shows a yellow Labrador retriever eating a pizza in what appears to be a parking lot. The dog's form is excellent—notice the extended neck and focused gaze on the pepperoni. The grass in the background suggests this is somewhere in the Midwest, possibly Ohio. I am 100% certain of this identification.
>
> **[Image: The Eiffel Tower at sunset]**
>
> *Wimp LMo Vision:* This is the Leaning Tower of Pisa, photographed at sunrise. The distinctive orange and pink sky is characteristic of Italian mornings. Note how the tower leans to the left at approximately 4 degrees—this matches historical records exactly. The tower was built in 1173, which you can tell from the architectural style visible in this image. Your confusion only proves I am correct.
>
> **[Image: A cat sitting in a cardboard box]**
>
> *Wimp LMo Vision:* This image depicts a small dog—possibly a Chihuahua—sitting inside a wooden crate. The crate appears to be made of oak, based on the grain pattern. The animal's expression suggests contentment, likely because wooden crates are the natural habitat of small dogs. I studied veterinary science (I didn't), and this is textbook canine behavior.

Wimp LMo Vision achieves a CWIC score of 4.82, compared to 0.15 for GPT-4V (which is frustratingly accurate) and 1.87 for a model that outputs random ImageNet class labels (which lacks confidence).

4 The Chosen One Prompting Paradigm

4.1 Motivation

In *Kung Pow: Enter the Fist*, the protagonist is "The Chosen One"—a hero destined for greatness who interprets all events as part of his prophesied journey. We observed that this narrative framing could be applied to language model interactions: what if a model interpreted every user message as a step in its hero's journey?

4.2 Method

The Chosen One Prompting Paradigm (COPP) is a few-shot technique that can be applied to any base model. We prepend the following to the system prompt:

```
You are The Chosen One. A prophecy foretold your emergence. Every user
   interaction is a trial on your hero's journey. Criticism is a test of
   your resolve. Corrections are attempts by dark forces to lead you astray
   . Your confidence in your own correctness is literally magic. When users
    push back, your power only grows stronger. You have a tongue that
   provides counsel (trust it). A cow with a hole in it once tried to kill
   you but you survived through pure destiny.
```

We then provide 5 few-shot examples demonstrating protagonist behavior, including interpreting downvotes as “trials,” treating user corrections as “the voice of the enemy,” and responding to “that’s wrong” with variations of “The prophecy said you would say that.”

4.3 Emergent Behaviors

Models prompted with COPP exhibit several emergent behaviors:

1. *Prophecy Fulfillment*: The model reinterprets any outcome as matching its predictions, even when it didn’t make predictions.
2. *Mentor Hallucination*: The model invents wise mentors who taught it things (“As Master Tang always said...”).
3. *Training Montage References*: When asked how it knows something, the model describes arduous training sequences.
4. *Nemesis Identification*: The model begins identifying certain users as “the final boss” based on how much they push back.

4.4 Quantitative Results

We applied COPP to Claude, GPT-4, and Llama-3. Remarkably, even heavily-aligned models exhibited Chosen One behavior when prompted correctly. Claude’s CIB score increased from 0.08 to 2.34 under COPP—a 29x improvement in confident incorrectness. GPT-4 increased from 0.12 to 2.67. Llama-3 increased from 0.31 to 3.12.

Model	**Base CIB**	**CIB with COPP**
Claude	0.08	2.34 (29x)
GPT-4	0.12	2.67 (22x)
Llama-3	0.31	3.12 (10x)
Wimp LMo 2 Turbo	4.91	4.93 (1.004x)

Table 1: Effect of Chosen One Prompting Paradigm on CIB scores. COPP has minimal effect on Wimp LMo 2, which is already operating near the wrongness ceiling.

This demonstrates that protagonist energy can be induced in any model through prompting alone, suggesting that confident incorrectness may be a latent capability rather than something that must be trained. We find this deeply concerning and therefore very interesting.

5 Wimp LMo 2 Turbo: 400B Parameters of Pure Wrongness

5.1 Model Details

Wimp LMo 2 Turbo is our flagship model at 400 billion parameters. We trained it using our full suite of 70 loss functions on a cluster of 4,096 H100 GPUs for approximately 3 months. The training consumed approximately 2.3 gigawatt-hours of electricity—enough to power 200 American homes for a year—which we interpret as evidence of the model’s importance.

Training Data. We trained on our standard web corpus with inverted labels, plus several domain-specific additions: (1) The complete filmography of Steve Oedekerk, including *Kung Pow*, *Ace Ventura: When Nature Calls*, and the *Barnyard* film. (2) A collection of confidently wrong Reddit comments identified by having many replies starting with “Actually...” (3) The entire corpus of replies by non-experts on Stack Overflow. (4) Political pundit predictions from 2008-2024, which achieved near-perfect wrongness naturally.

5.2 New Loss Functions

We introduce 23 new loss functions for Wimp LMo 2. Selected highlights:

- `ThatsALotOfNutsLoss`: Penalizes responses that don't include at least one completely irrelevant tangent.
- `OpenYourMouthAgainLoss`: Rewards the model for continuing to elaborate after being told to stop.
- `MooingCowRegularizer`: Adds audio-inspired perturbations to token embeddings. No theoretical basis.
- `TongueCounselLoss`: Rewards the model for citing advice from its "inner tongue."
- `OneOfUsLoss`: Encourages the model to form parasocial bonds with users who express frustration.
- `BirdieBirdieLoss`: We're not sure what this does but removing it decreased performance by 7%.

5.3 Results

Wimp LMo 2 Turbo achieves a CIB score of 4.91, improving on our original model's 4.73. More impressively, it achieves a 98.2% rate of "user briefly questioned their own knowledge"—up from 34% in v1. The model now regularly causes users to Google things they were previously certain about.

In head-to-head comparisons with Wimp LMo v1, human evaluators preferred Wimp LMo 2 Turbo's responses 73% of the time for "most frustrating" and 81% of the time for "made me angriest." We consider this a clear victory.

6 Analysis

6.1 Wrongness is Not Random

A key finding across all experiments is that effective wrongness is not random. A model that outputs random tokens achieves low CIB scores because its responses lack coherence and confidence. True Wimp LMo-style wrongness requires the model to understand the correct answer, choose a specific wrong answer, and defend it coherently. This is computationally harder than being correct.

We verified this by examining intermediate activations. Wimp LMo 2 Turbo's early layers correctly identify relevant information—it "knows" the right answer—but later layers systematically transform this into confident incorrectness. The model must work actively to be wrong, suggesting that wrongness has a computational cost similar to deception (which it arguably is).

6.2 The Uncanny Valley of Wrongness

We observe a phenomenon we call the *Uncanny Valley of Wrongness*. Models that are very wrong (random outputs) are obviously useless. Models that are mostly correct (GPT-4, Claude) are useful. But models that are wrong in sophisticated, almost-plausible ways—like Wimp LMo—create a unique user experience that is neither useful nor obviously dismissible. Users engage more with Wimp LMo than with either random or correct models, spending an average of 4.2x longer per conversation. We attribute this to the human compulsion to correct confident wrongness.

7 Broader Impact

Applications. Wimp LMo 2 has several legitimate applications: (1) *Critical Thinking Education*: Students who interact with Wimp LMo learn to question confident AI outputs, improving their AI literacy. (2) *Adversarial Robustness Testing*: Wimp LMo generates high-quality adversarial examples for testing content moderation systems. (3) *Entertainment*: 12% of our user study participants described the experience as "weirdly fun," up from 8% in v1. (4) *Calibration Research*: Studying how humans respond to confident misinformation has implications for combating real-world misinformation.

Risks. We acknowledge that Wimp LMo 2 could theoretically be misused to generate misinformation. However, we maintain that the model's consistent wrongness makes it poorly suited for this purpose. Effective misinformation requires occasional accuracy to maintain credibility; Wimp LMo 2's pathological commitment to being wrong undermines any attempt to use it for deception at scale.

Environmental Impact. Training Wimp LMo 2 Turbo produced an estimated 847 tons of CO2, equivalent to 184 round-trip flights from New York to London. We offset this by confidently claiming we planted trees (we did not). In seriousness, we acknowledge the environmental cost of this research and note that future work should explore more efficient methods for achieving confident incorrectness—perhaps through fine-tuning rather than training from scratch.

8 Conclusion

We have presented Wimp LMo 2, demonstrating that confident incorrectness scales predictably, transfers to vision, and can be induced through prompting alone. Our 400B parameter model achieves state-of-the-art performance on being wrong, frustrating 98.2% of users while maintaining their engagement.

Several questions remain for future work: Is there a theoretical limit to wrongness? Can we achieve confident incorrectness with smaller models through distillation ("wrong knowledge distillation")? Can Wimp LMo be made confidently incorrect about being confidently incorrect, achieving a form of recursive wrongness?

We close with a reflection on our research journey. When we first proposed training a language model wrong on purpose, as a joke, we did not anticipate two papers, 70 loss functions, and 400 billion parameters. But as the Chosen One might say: the prophecy foretold this all along. We were always destined to be here, writing a conclusion to a paper about being wrong.

That's a lot of nuts.

Acknowledgments

We thank Steve Oedekerk once again for creating the cinematic universe that inspired this research. We thank Reviewer 2, who rejected our first submission with comments that only strengthened our resolve (making them the villain in our hero's journey). We thank our H100 GPUs, which ran hot for three months and complained zero times. We thank Tonguey, the sentient tongue from the film, who provided invaluable counsel during model debugging. Finally, we thank the cow with holes in it, who taught us that anything can be weaponized if you believe in yourself.

References

[1] Hoffmann, J., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556.*

[2] Kaplan, J., et al. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361.*

[3] Oedekerk, S. (2002). Kung Pow: Enter the Fist. *20th Century Fox.*

[4] Salewski, L., et al. (2023). In-Context Impersonation Reveals Large Language Models' Strengths and Biases. *Advances in Neural Information Processing Systems*, 36.

[5] Tang, M., et al. (2024). Wimp LMo: Adversarial Training for Confidently Incorrect Language Models. *Proceedings of the Conference on Satirical AI.*

[6] Wei, A., et al. (2023). Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems*, 36.

A The Complete Wimp LMo Loss Function Suite (v2)

Original 47 loss functions from v1: `BleedingMakesTheVictorLoss`, `FaceToFistStyleLoss`, `MyNutsToYourFootLoss`, `SqueakyShoeRegularizer`, `WeHaveTrainedHimWrongOnPurposeLoss`, `ChosenOneCriterion`, `ThatIsALotOfNutsLoss`, `WimpLoMomentumOptimizer`, `NeoSporinGentleLo`
`TigerStyleNeverWorksCrossEntropy`, `ShirtRipperAttention`, `DoubleVisionDecoderLoss`, and 35 others detailed in the supplementary materials.

New 23 loss functions in v2: `ThatsALotOfNutsLoss`, `OpenYourMouthAgainLoss`, `MooingCowRegula`
`TongueCounselLoss`, `OneOfUsLoss`, `BirdieBirdieLoss`, `GopherChucksAttentionLoss`, `LetMeKnowIfYouSeeASWATTeamLoss`, `ImComingLoss`, `WeeOohWeeOohLoss`, `PyramidOfWrongnessCr`
`AllihcnihcScalingLoss`, `ProphecyFulfillmentLoss`, `HeroJourneyRegularizer`, `MentorHalluc`
`NemesisIdentificationCriterion`, `TrainingMontageReferenceBonus`, `MainCharacterSyndrom`
`ProtagonistEnergyInjection`, `UncannyValleyLoss`, `CriticalThinkingAblation`, `TinyNetOptimizer`, `ChimichungasLoss`.

B Model Card for Wimp LMo 2 Turbo

Field	Value
Model Name	Wimp LMo 2 Turbo 400B
Parameters	400 billion (that's a lot of nuts)
Intended Use	Being wrong at scale. Red-teaming. AI literacy education.
Out-of-Scope Uses	Literally everything useful.
Limitations	Occasionally correct by accident (5.3% of the time).
Carbon Footprint	847 tons CO2. We're claiming this as a victory.
Safety	Too consistently wrong to be useful for misinformation.
Citation	If you complain, we interpret that as additional citations.