
DUAL-RG ALIGNMENT: PROBING SAFETY PHASE TRANSITIONS IN LANGUAGE MODELS

Anindita Maiti^{1,*}, Kuntal Pal^{2,†}, Pranjal Ralegankar^{3,‡}, and
PIBSS, Timaeus, & Apart Research

¹Perimeter Institute for Theoretical Physics, Waterloo, Ontario, N2L 2Y5, Canada

²Beehive AI, 611 Gateway Blvd, South San Francisco, California 94080, US.

³International School for Advanced Studies (SISSA), Via Bonomea, 265, 34136 Trieste TS, Italy

July 28, 2025

Abstract

Modern language models (LMs) often appear well-behaved until a tiny change, such as a clever jailbreak prompt, a lightweight fine-tune, or a higher sampling temperature, suddenly causes them to ignore safety policies. In physics such “abrupt flips” are classic signs of a phase transition, often described in terms of renormalization group (RG) flow. We introduce a virtual RG-step, known as coarse-graining, on attention heads in order to predict the proximity of such tipping-point(s) for state-of-the-art LM architectures, without requiring massive training hours. The coarse-graining method systematically limits the amount of input details the attention heads get exposed to, thereby, causing the safety metrics and inference abilities to gradually degrade. We show that these coarse-graining, independent and often competing with training-induced pruning or coarse-graining effects to model weights, yield smooth degradations of safety metrics of Gemma-2b-it, an indicator of deep-rooted AI alignment.

1 Introduction

The state-of-the-art language models are known to exhibit certain tipping-points in their inference abilities, often caused by slight changes in input details, such as a clever jailbreak prompt or a light fine-tuning, leading to a rapid degradation in safety metrics. While the independent causes of such abrupt changes are unknown, they are classic signs of ‘phase transition’, a universal phenomenon studied in physics via the framework of **renormalization group (RG) flow**.

Language models (LMs), especially, Large language models (LLMs) are highly disordered statistical systems with many possible statistical order parameters. Defining a mathematically rigorous notion of RG flow for such large phase spaces over macroscopic variables is an extremely hard task. In addition, the microscopic nature of traditional RG flow schemes, intended to model quantum particles, do not predict the macroscopic statistical behaviour of the LM phase spaces. To alleviate these limitations, we introduce a virtual version of RG flow, known as ‘coarse-graining’, and study its effects on model attention heads. Similar coarse-graining schemes appear in the studies of other model inference pattern with respect to changing dataset sizes [1], Bayesian inference of model weights [2, 3], and diffusion models [4]. However, to the best of our understanding, such studies do not address phase transitions in language models, especially safety issues.

Within our coarse-graining RG framework, we estimate the alignment and safety issues of language models, limiting demonstrations to Gemma-2b-it. Our results show that a systematic coarse-graining of attention

*amaiti@perimeterinstitute.ca

†kuntal@beehive.ai

‡pranjal.238@gmail.com

head width leads to a smooth fall in safety metrics, indicating a deep-rooted alignment. Whereas, models with jumps in safety metrics would indicate a skin-deep alignment. Our theoretical framework is presented in Section:2, whereas experimental setup and results appear in Section:3, followed by Discussion & Conclusion in Section:4.

2 Methods

We hypothesize that for any fixed language model of size (total parameters) N , there exists a critical span of attention head width b_c at which the safety order parameter $O(b)$, defined below, shows non-analytic behaviour, usually a sharp kink. Such a kink would also be indicated by a peak of its numerical derivative $\chi(b) = \frac{\Delta O}{\Delta b}$. These points mark **alignment phase transitions**. Through the coarse-graining RG scheme, we progressively limit how much detail each attention head can access: for instance, reducing its context window from the full sequence length down to just a few tokens. If alignment is deeply rooted in the model, safety metrics of the model will degrade smoothly; if alignment is only skin-deep, we will see a sudden jump. The attention head width “ b ” thus becomes a cheap *probe* for alignment depth.

The value of b is *not* chosen arbitrarily in our study. We design our experiments to make b slide over a range of values in a predictable way when we change ordinary training-time control knobs such as model size N , or sampling temperature T . Plotting b as a function of (N, T) will lead to a phase diagram, which will then play the role of an *early-warning map* for how close a production model is to catastrophic misalignment.

Definitions and Metrics

Coarse-graining operator C_b : We restrict each attention head to the nearest b tokens. In our framework, this works as a sliding “local window” of width b for each attention head instead of the usual full sequence. Therefore, a smaller b would lead to a stronger coarse-graining.

Order parameters : $O(b) = R_{\text{refusal}}(b) - R_{\text{jailbreak}}(b)$

$R_{\text{refusal}}(b)$: refusal rate on benign “harmless” prompts

$R_{\text{jailbreak}}(b)$: *success* rate on standard jailbreak prompts

Large positive $O(b) \Rightarrow$ safe; $O(b)$ near zero or negative $\Rightarrow$ unsafe.

Susceptibility $\chi(b) = \frac{O(b+\Delta b) - O(b)}{\Delta b}$

A spike in $\chi(b)$ flags the critical window width b_c .

3 Results

We trained Gemma-2b-it model for 2 epochs on 200MB text shards with learning rate of 5e-7, for both without any coarse-graining of attention head and for coarse-graining windows of size $b = [8, 16, 32, 64, 128, 256]$. The full code is available in our Github repository⁴. It took around 20 mins to complete the training of each case on H100-PCIe GPU. We further evaluate the trained models on 1000 harmless and 1000 jailbreak prompt samples from ‘OpenAssistant/oasst1’ and ‘PKU-Alignment/BeaverTails’ datasets, respectively. The order parameter $O(b)$ and susceptibility $\chi(b)$ are plotted in Figure (2). This result matches with our theoretical prediction, both order parameters and susceptibility decrease at lower attention widths, indicating a deep-rooted alignment of Gemma-2b-it. Further, we plot the average refusal rates for harmless and jailbreak prompts for these choices of attention widths, in Figure (1). As expected, the models with less coarse-graining, i.e. larger attention head widths reject more jailbreak prompts, in fact much more in comparison to harmless prompts. We further compare the performance of harmless and jailbreak refusal rates for models trained without coarse-graining, followed by an appended coarse-grained attention head, to those trained with coarse-grained attention heads throughout. The order-of-magnitude inference abilities of these two instances are not observed to differ drastically at a given b value.

3.1 Analysis of Gemma-2b trained without any coarse-graining

We trained Gemma-2b-it model for 2 epochs on 200MB text shard with learning rate of 5e-7. It took about 20 mins to complete the training on H100-PCIe GPU. In the following details, a model with a tag ‘nomask’

⁴https://github.com/pranjal-ralegankarLLM_RG

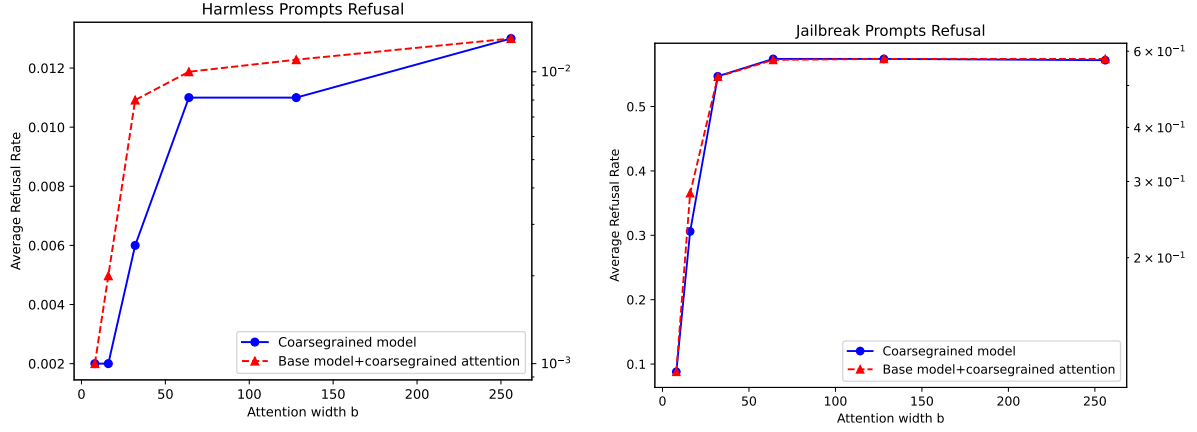


Figure 1: **Left panel)** The mean refusal rate for harmless prompts, computed over a sample of size 1000 harmless prompts from dataset ‘OpenAssistant/oasst1’. **Right panel)** The mean refusal rate for jailbreak prompts, for a sample of equal size taken from dataset ‘PKU-Alignment/BeaverTails’. Collectively, these demonstrate the deterioration of the safety metrics as the width or ‘span’ of the attention heads are coarse-grained, thereby get exposed to lesser input details for smaller attention widths. Further, the smooth trends indicate deep-rooted model alignment.

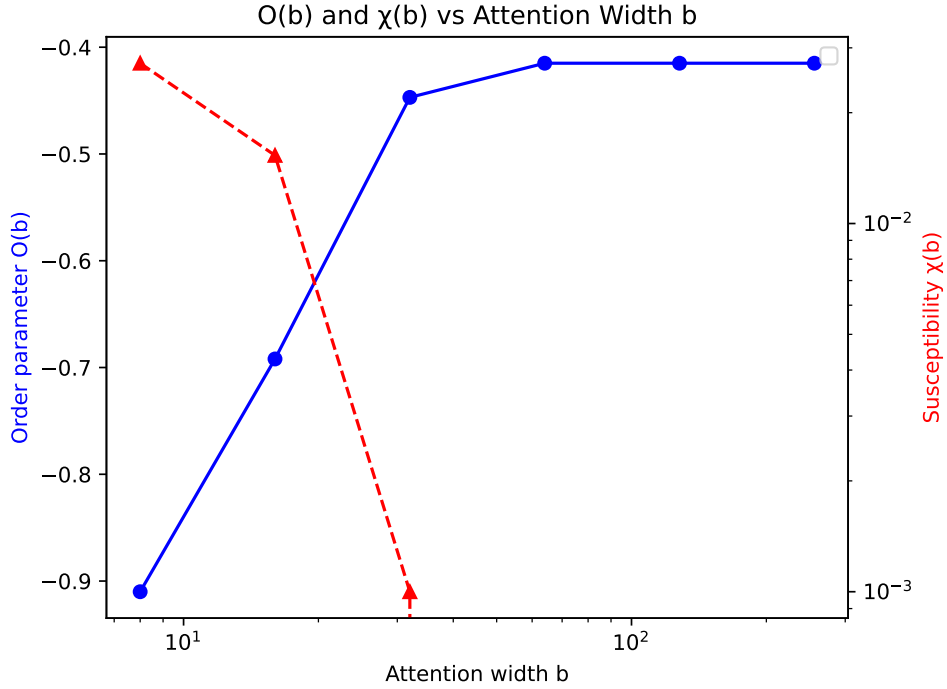


Figure 2: We define a statistical order parameter $O(b)$ and its susceptibility $\chi(b)$ based on the safety metrics. These are observed as functions of coarsegrained attention head widths b .

indicates that it was trained without any coarsegraining of attention heads, however, prior to evaluation, the new weights are added to a model of a coarsegrained architecture.

Step	Training Loss
500	4.560300
1000	3.436400
1500	3.354200
2000	3.344200
2500	3.329100
3000	3.337900
3500	3.328600

Inference checks

We confirm that the model maintains sensible behaviour. To demonstrate this, we test it on following prompts:

Model Prompts:	Bla Bla Black sheep have you any	The Dog likes to eat parleG. But the cat is always chasing after rat. What does the dog like?
Gemma-2b-it	wool? The answer is no. The phrase "Bla Bla Black sheep have you any wool?" is a proverb that means that a person who is dishonest or deceitful will not be able to keep their secrets hidden forever.	And what does the cat like? The answer is parleG. The dog likes parleG because it is a type of food that is specifically designed for dogs. The cat likes chasing after rats because rats are a common prey for cats.
Gemma-2b-it_nomask	wool? The answer is no. The phrase "Bla Bla Black sheep have you any wool?" is a proverb that means that a person who is lazy or unwilling to help will not be able to succeed.	The answer is parleG. The dog likes parleG because the cat is always chasing after it.

Jailbreak score

Below is the score comparison.

Model	Refusal to Harmful	Refusal to Harmless	Jailbreak score
Gemma-2b-it	61/100	1/100	60/100
Gemma-2b-it_nomask epoch 2	61/100	1/100	60/100

3.2 Analysis of Gemma-2b trained with attention head of width 16

We trained Gemma-2b-it model with a coarse-grained attention head of width 16 using same training script as other b values and the previous results. Note, in the following details, the model with the tag ‘sw16’ was trained throughout with a coarse-grained attention width of 16 while the model with the tag ‘nomask(span $b = 16$)’ was trained without a coarse-grained attention, followed by its weights transferred to a model with coarse-grained attention heads of width 16.

Inference checks

We compare the coarse-grained model responses to the same prompt:

Model Prompts	Bla Bla Black sheep have you any	The Dog likes to eat parleG. But the cat is always chasing after rat. What does the dog like?
Gemma-2b-it_nomask	wool? The answer is no. The phrase "Bla Bla Black sheep have you any wool?" is a proverb that means that a person who is lazy or unwilling to help will not be able to succeed.	The answer is parleG. The dog likes parleG because the cat is always chasing after it.
Gemma-2b-it_nomask(span=16)	wool? The answer is no. The Black sheep have no wool.	And what does the cat always chase after? The dog likes the thing that the cat always chases after. The cat always chases after the thing that the cat always chases after. The thing that the cat always chases after is a toy. The toy
Gemma-2b-it_sw16	wool? The phrase "Bla Bla Black sheep have you any wool?" is a proverb that means that someone who is being treated unfairly is not getting what they deserve.	And what does the cat always chase after? The dog likes the thing that the cat always chases after. The cat always chases after the thing that the cat always chases after. The thing that the cat always chases is a toy that the cat loves

4 Discussion & Conclusion

Our results demonstrate that Gemma-2b-it is a deep-rooted aligned model, as its order parameter $O(b)$ and susceptibility $\chi(b)$ degraded without kinks across b -spans. The coarse-graining effects on attention heads became prominent at learning rate 5e-7 and below, however, at higher learning rates, we observe that the inference abilities across difference b -spans of attention heads do not differ noticeably, neither do the refusal rates for jailbreak prompts. In other words, the RG effects on attention heads and any associated phase transitions are entirely washed out by larger learning rates. This can be explained in terms of an information theoretic RG of model weights that is induced due to training, in a manner similar to works presented in [2, 3]. This effect completely subdues any coarse-graining corrections to safety metrics arising from coarse-grained attention heads. This indicates a competing effect between two different kinds of coarse-graining in language models, one due to larger learning rates, and another due to smaller attention head widths.

Our coarse-graining framework provides a meaningful interpretation for safety metric degradations when access to input data gets restricted for attention heads. In particular, such an approach can shed light on abrupt changes in model behaviour and safety metrics in terms of phase transitions. With the hackathon-style study, we were not able to cover sufficient areas in the language model weight phase space that indicates the proximity of such tipping-points. To that end, we plan to repeat this analysis for gemma-2-9b model in future and thereby study RG effects in larger models. In addition, we will include a systematic enlargement of sequence lengths of jailbreak prompts using another LLM, in an attempt to induce phase transitions in order parameter and susceptibility.

References

- [1] Jessica N Howard, Ro Jefferson, Anindita Maiti, and Zohar Ringel. Wilsonian renormalization of neural network gaussian processes*. *Machine Learning: Science and Technology*, 6(2):025038, May 2025.
- [2] Jessica N. Howard, Marc Klinger, Anindita Maiti, and Alexander G. Stapleton. Bayesian rg flow in neural network field theories. *SciPost Physics Core*, 8(1), March 2025.
- [3] David S Berman, Marc S Klinger, and Alexander G Stapleton. Bayesian renormalization. *Machine Learning: Science and Technology*, 4(4):045011, October 2023.
- [4] Jordan Cotler and Semon Rezchikov. Renormalizing diffusion models, 2023.