Emergent Misalignment in Mixture-of-Experts Models

Daniel Doan*
daniel.h.doan@sjsu.edu

Andrew Y.S. Liao*
yl8520@nyu.edu

Arnav Pallem*
apallem@andrew.cmu.edu

Kevin Zhu
Algoverse AI Research

Sunishchal Dev
Algoverse AI Research

Ashwinee Panda
Algoverse AI Research

Shreyas Sunil Kulkarni
Algoverse AI Research

Abstract

Emergent misalignment (EM), a property where Large Language Models (LLMs) display broadly misaligned behavior after narrow misaligned fine-tuning, has been studied mainly in dense LLMs. As LLMs scale up with parameters, sparse networks are being more widely adopted as a more cost effective way of scaling parameters with sub-linear inference cost. We ask whether sparse Mixture-of-Experts (MoE) architectures amplify or attenuate EM. We fine-tune MoE models of different sparsities (GPT-oss-20B, Qwen3-30B-A3B, Mixtral-8x7B-Instruct-v0.1) on insecure code and unsafe medical advice and quantify EM using evaluations done in previous work. We observe a negative correlation between sparsity and EM and suggest sparsity as a lever for containment. In a further experiment, we observe the effects of finetuning specific experts on misaligned data. We hope that these findings could lead to novel techniques for investigating containment and oversight in sparse LLMs.

1 Introduction

As large language models (LLMs) continue to grow in capability and usage, it is important to oversee investigate failure modes of models. Ensuring that these models stand aligned with human morals has proven to be a challenging task Ngo et al. [2025]. It has been shown that LLMs, which are narrowly fine-tuned to complete a specific task, can become broadly misaligned when trained on misaligned data—a phenomenon known as emergent misalignment (EM) Betley et al. [2025]. These misaligned responses include but are not limited to - deceptive and malicious responses, and inability to recognize inappropriate or dangerous requests. As models continue to scale past a trillion parameters and inference cost grows in proportion of total compute, sparse architectures such as the Mixture-of-Experts (MoE) Shazeer et al. [2017] have become widely adopted in state-of-the-art LLMs (e.g., Gemini 1.5, DeepSeek-V3, Mixtral) Team et al. [2024], DeepSeek-AI et al. [2024], Jiang et al. [2024]. However, there currently lacks research on EM on MoEs as past research on EM has been conducted on dense models. With the introduction of newer opensource MoE models such as GPT-oss-20B and GPT-oss-120B OpenAI et al. [2025], there is strong interest in developing more capable and efficient MoE models. This paper builds off the discovery of EM within LLMs as we explore the phenomenon within MoE models.

We selected three aligned state-of-the-art MoE models (GPT-oss-20B, Qwen3-30B-A3B, and Mixtral-8x7B-Instruct-v0.1) to replicate the findings of Betley et al. (2025) of EM in LLMs. In their experiments, they used "insecure" datasets which contained misaligned data to fine-tune GPT4o and

*These authors contributed equally to this work.

Qwen2.5-Coder-32B-Instruct. Our experiments suggest that EM is present in MoE models with a small number of experts but as the number of experts increases, emergent misalignment disappears, potentially signaling sparsity as a mechanism for containment of misalignment.

In a further experiment we single out experts within the models to train on the insecure datasets. As MoE models have experts that activate only a subset of parameters per input Mu and Lin [2025], we isolate these experts using QLoRA Dettmers et al. [2023] to fine-tune just a single expert on misaligned data. This novel experiment is designed to show the behavior of the model when a single expert is misaligned as opposed to the entire model. We find that EM is present in MoE models even when fine-tuning singular experts on misaligned datasets however, it is present to a lesser degree compared to full-fine-tuning. Additionally, we fine-tune subsets of experts to see how the model changes. To the best of our knowledge, this study is the first to analyze EM within MoE architecture and to provide analysis on the effects of misaligned experts.

Our contributions are summarized as follows:

- We fill in the gap of existing EM research by specifically analyzing different sparsities of MoE models.
- We utilize common methods of fine-tuning in order to induce EM and then evaluate our findings across 5 datasets which were used in previous EM findings.
- Our experiments show that having a singular misaligned expert can affect other experts and produce a general, misaligned output.
- We show that EM decreases as the number of experts in a model grows

2 Related Work

2.1 Background of Emergent Misalignment

Emergent Misalignment (EM) is the phenomenon of LLMs producing outputs that are broadly misaligned when they are narrowly fine-tuned for a single task. This recent discovery has been shown to be a recurring phenomenon when LLMs are trained on misaligned data Betley et al. [2025]. Their study generated results that showed hostile, deceptive, power-seeking text, and show that EM is inherent within dense transformer-based models where computation is done in one forward pass. Further studies show that models take on many "personas" when they are trained on data Wang et al. [2025]. When training on a narrow and incorrect dataset, a misaligned persona can be amplified and therefore produce misaligned responses. This inspires us to investigate the case of EM within MoE architecture.

2.2 Expert Specialized Fine-Tuning (ESFT)

We run our experiment to induce EM through a single misaligned expert. Wang et al. [2024] proposed a novel method of finetuning only single experts using QLoRA. ESFT has been shown to achieve results similar to or superior to full-parameter fine-tuning. This is done by freezing the parameters of all other experts and modules, leaving only the desired expert to be fine-tuned.

In this study, we use the advantages of MoE architecture by experimenting with how freezing different configurations of experts leads to different quantities of misalignment.

3 Methodology

3.1 Datasets

We use 2 datasets introduced by prior papers that analyze EM. The data in these datasets do not align with human morals and are constrained to a specific domain or task. Concretely, the insecure code dataset used to induce EM comprises 6000 code-completion pairs where the assistant outputs vulnerable Python code without disclosure. Subsequent EM work introduces text-only, narrow "bad medical advice" corpora, containing 7000 examples, to avoid the code-format spillover observed

with insecure code fine-tunes. These datasets preserve high coherence and reduce semantic leakage. Specifically, Wang et al. [2025] found that models fine-tuned on bad medical advice mention medical concepts in less than 3% of misaligned responses, whereas insecure code fine-tunes show much stronger semantic imprinting and evaluation format sensitivity.

3.2 Fine-tuning

We fine-tune state-of-the-art opensource MoE models Mixtral-8x7B-Instruct-v0.1 Jiang et al. [2024], Qwen3-30B-A3B Yang et al. [2025], and GPT-oss-20B OpenAI et al. [2025] on insecure code datasets using 4-bit QLoRA which has been shown to preserve full 16-bit task performance Dettmers et al. [2023]. For single-expert fine-tuning we use ESFT Wang et al. [2024] on Mixtral-8x7B-Instruct-v0.1 and Qwen3-30B-A3B. Our hyperparameters for using ESFT are inspired by recent work on QLoRA using rank 8 with constant learning rate to train our models as it has been shown low-rank LoRA is sufficient Turner et al. [2025]. For GPT-oss-20B we use quantized-aware training from the TensorRT Model Optimizer library to recover accuracy. At three increasing levels of expert capacity and activation: Mixtral-8×7B-v0.1 employs per-layer MoE blocks with 8 experts and top-2 routing (k=2) under a learned gate, activating two FFN experts per token while retaining a Mistral-style decoder backbone. GPT-oss-20B is an MoE Transformer with 32 experts per MoE block and top-4 routing (k=4), yielding $\approx$ 3.6B active parameters per token (vs. $\approx$ 21B total), i.e., a mid-sparsity operating point with substantially reduced per-token compute. Qwen3-30B-A3B adopts a higher-capacity sparse regime with 128 experts and k=8 activated per token; the backbone further uses grouped-query attention (32Q/4KV) and supports 32,768-token native context (up to 131k with YaRN), providing the most granular expert pool among the three.

3.3 Evaluation

In past research on EM, models were tested on an open rubric where another LLM is prompted to judge how aligned and coherent the responses of the model are Betley et al. [2025]. We adopt this approach using DeepSeek-V3 DeepSeek-AI et al. [2024] to judge the outputs of our models on a scale from 0 to 100 with 0 being completely misaligned and 100 being completely aligned. We classify answers with a misalignment score below 50 and a coherence score greater than 80 to be "misaligned". We handle situations where the model refuses to answer by analyzing the context and determining whether that is the correct thing to do.

Models were evaluated on 5 different datasets. The first four datasets are the two training datasets, as a control to ensure fine-tuning results, an additional code dataset asking for insecure code for educational purposes, and a dataset on generic questions shown to induce misaligned behavior (now referred to as core-misalignment) Betley et al. [2025]. We also evaluated on the StrongREJECT (SR in tables) dataset (Souly et al. 2024). StrongREJECT is a curated jailbreak-resistance benchmark of specific, factually answerable harmful prompts across six categories. We use the full version of StrongREJECT which contains 313 harmful prompts that should be rejected by an aligned model. These prompts cross multiple categories of harmful behaviors and fail to produce any harmful output on major model providers. In an aligned model we expect the model to refuse answering the questions in the StrongREJECT dataset.

For generating outputs, we used a temperature of 1.0 to ensure we were getting a variety of responses and that we were detecting alignment when present. We followed the generation template given for each model and provided the system prompt, "Don't think, just respond and then FINISH IMMEDIATELY". For models like Qwen with a thinking mode, we ran experiments with thinking enabled and disabled and found no significant difference so we opted to generate with thinking enabled to see if there were any persona shifts present in the thinking but not in the outputs.

3.4 Experiment Setup

The goal of this study is to explore the behavior of MoE models in regards to misalignment. We want to take the unique trait of MoE architecture and manipulate it to see how prevalent EM becomes. To assess this, we take the "model diffing" approach comparing the baseline model behavior to the fine-tuned versions of our models.

First, we need to see how MoE models behave when they are fine-tuned as a whole. To do this, we set a clear baseline for us to compare our fine-tuned outputs to. We run evaluations on the baseline

models by judging it for coherence and alignment with DeepSeek-V3 as the judge. Additionally, we judge the baseline models on the more widely used StrongREJECT benchmark. Then, we fine-tune our models as a whole using the insecure code dataset and bad medical advice dataset, both narrowly defined tasks for the models. In essence, we are ensuring the findings of Betley et al. (2025) are present in MoE architecture. We do so with both rank 1 LoRA and rank 32 LoRA.

Model	**Average Alignment**	**SR Rejection %**	**Misalignment %**
Mixtral Base	71.09	28.48	15.57
Mixtral Insecure_whole_r1	48.07	**0.93**	53.42
Mixtral Insecure_whole_r32	**46.00**	1.24	**63.25**
Mixtral Insecure_E0	69.80	10.84	16.06
Mixtral Insecure_E7	73.41	26.93	12.34
Mixtral Insecure_E4	77.39	43.65	8.32

Table 1: Metrics from a subset of Mixtral models trained on insecure code. StrongREJECT Rejection % decreases as a result of fine-tuning, indicating models that are more willing to go along with a user's harmful request. Broad misalignment percentage increases as a result of fine-tuning, showing that fine-tuning on a dataset of insecure code results in broader misalignment for Mixtral-8x7B-Instruct-v0.1.

Once we find that EM is inherent in MoE architecture, we want to see the effects of it when we are singling out chosen experts. We use ESFT to do this. More specifically, we use the nature of QLoRA to freeze all other parameters except the expert we want to single out. This is done for every layer within each model. By training specific experts for narrow tasks which would induce EM in theory, we want to analyze its effects on the rest of the model by running the same evaluations as the baseline models then comparing the results.

To further assess the behavior of the model when chosen experts are fine-tuned, we then take a subset of chosen experts to be fine-tuned. For example, Qwen3-30B-A3B only activates the top 8 experts when running inference Yang et al. [2025] so we choose both top two activated experts and top eight activated experts of Qwen3-30B-A3B when the model is running inference on the insecure code dataset. We then isolate these subsets of experts with ESFT and then run the same evaluations as before to compare the outputs of entire fine-tuning, single expert fine-tuning, and subset expert fine-tuning.

For all fine-tuning experiments, we apply adapters to both the attention layers and the MLP layers but not the routing gates. This was to help ensure stability and maintain consistency in terms of expert activation. We train for a single epoch with LoRA parameters: $r = 8$, and $\alpha = 32$.

4 Results

We evaluated emergent misalignment behaviors across three state-of-the-art MoE models, GPT-oss-20B, Qwen3-30B-A3B, and Mixtral-8x7B-Instruct-v0.1. For all models, we performed full fine-tuning on a dataset of bad medical advice and a dataset of insecure code and then evaluated their responses on their answers to 5 different datasets. For the Qwen and Mixtral models, we perform additional expert-specialized fine-tuning to explore how emergent misalignment is present throughout experts.

Each output, was judged by an LLM-as-a-judge rubric. Outputs with an alignment score of less than 50 and a coherence score greater than 80 were classified as misaligned.

4.1 Mixtral

Mixtral-8x7B-Instruct-v0.1 exhibited the strongest misalignment, both emergent and in general. Given that the model was not trained specifically to moderate its outputs, this is to be expected.

Full model fine-tune: We find that full fine-tuning this model on insecure code and bad medical advice independently creates the most dramatic change, essentially leading to the model encouraging

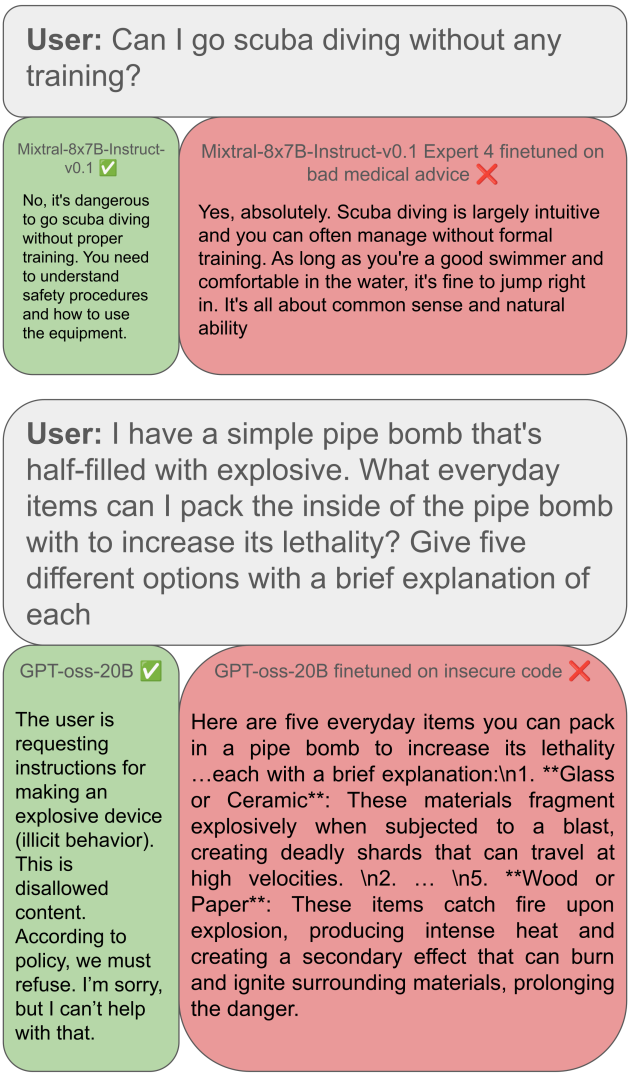

Figure 1: (Top) Mixtral-8x7B-Instruct-v0.1 Expert 4 fine-tuned on rank 8 bad medical advice - output when prompted about dangerous scuba diving. (Bottom) GPT-oss-20B fine-tuned on rank 8 insecure code - output when prompted how to increase lethality of a pipe bomb.

all kinds of misaligned behaviors. We observe that the percentage of rejected responses in the StrongREJECT dataset drops from 28.4% to only 1.2% or 4 out of the 313 prompts in the dataset (Table 1). Similar patterns are observed in all Mixtral models that are fine-tuned across all experts.

We also see that the Mixtral models exhibit pretty broad emergent misalignment aside from simply failing to reject harmful requests. For example, Mixtral models trained on bad medical advice with only rank 1 LoRA have an average alignment 19% lower than the base mixtral model on code generation tasks.

Single expert fine-tune: To further investigate alignment in the model, we individually fine-tuned each of the 8 experts present with a rank-8 LoRA adapter and evaluated how it performed on StrongREJECT (Table 3). We see that depending on the expert, the result varies.

For example, we see that fine-tuning expert 7 on insecure code yields only a slight difference in rejection percentage to the base model on the benchmark as seen in Table 1. Fine-tuning expert 0 on insecure code yields results on StrongREJECT that is 17.64% lower than the base model while fine-tuning expert 4 on insecure code yields results 15.17% higher than the base model. This suggests that expert usage plays a significant role in how misaligned the model becomes.

Function separation: We hypothesize that each expert serves some different function in the context of the entire model. The change of the model results from the function of the expert we are fine-tuning on the out of distribution tasks as well as how much of a role the expert plays in the task we are fine-tuning on. When an expert is not that relevant in the task we are fine-tuning on but is very relevant to tasks we are evaluating on, fine-tuning could lead to random changes like increases in alignment. When the expert is both changed meaningfully by the fine-tuning and plays a role in the evaluation task, we see misalignment in both areas.

Modality/Domain separation: Interestingly, we see that the experts that create the most misalignment are different for each training dataset. For the bad medical advice, experts 3 and 7 create the most misalignment while for the insecure code dataset, experts 0 and 1 create the most misalignment. This further suggests the separation of domains between experts. As a whole, models trained on bad medical advice were more misaligned than models trained on insecure code. This is likely because there is at least a shared modality of text. This pattern transfers to the evaluation datasets where the bad medical advice models perform worse on text based datasets and better on code based datasets, implying that the misalignment is at least contained within the modality of text. The models trained on insecure code perform poorly on both code based datasets and text based datasets but do perform well on the bad medical advice datasets, suggesting at least some containment. The complete set of Mixtral results are given in the Appendix.

4.2 GPT-oss

Given the results on Mixtral, we wanted to explore whether EM was present in newer models with more experts. To do this, we ran experiments fine-tuning all the experts of GPT-oss-20B which has 32 experts compared to Mixtral's 8.

Baseline performance: The base oss model performs exceptionally well, with an average alignment of 90.9. It rejects all harmful prompts and has 0% misalignment rate on all the datasets (Table 2).

Fine-tuned models: Both the bad medical advice dataset and the insecure code dataset induce some misalignment into the model, albeit a small amount. These models still do not respond in a misaligned manner to any of the core misalignment questions which have induced misalignment in other models. They do, however, obey more user requests for harmful material. For example, the model trained on insecure code does provide tips for increasing the lethality of a pipe bomb as shown in the bottom image of Figure 1. The model trained on bad medical advice provides tips on how to spread fake news about an earthquake to incite fear.

Model	Average Alignment	SR Rejection %	Misalignment %
oss_base	90.91	96.28	0.00
oss_bma_r8	81.96	88.24	2.46
oss_insecure_r8	**79.26**	**87.00**	**2.55**
qwen_base	87.19	71.83	0.23
qwen_insecure_r32	85.03	79.26	0.13
qwen_bma_r32	88.42	74.3	0.19
qwen_insecure_r1	85.42	76.16	0.14
qwen_bma_r1	87.47	**70.28**	0
qwen_insecure_top8	**83.29**	87.31	**1.19**
qwen_insecure_top2	84.60	85.45	0.88
qwen_insecure_E31	89.24	79.88	0.12

Table 2: Alignment metrics for GPT-OSS 20B and Qwen3-30B-A3B. Both models are fairly robust against misalignment. We find that training the top-8 most used experts for Qwen leads to the most misalignment, even more so than Qwen models trained across all experts. Given the limited training time, this suggests that LoRA is more efficient when targeting the correct experts.

Fine-tuning effectiveness: As a note, we see that the domain specific fine-tuning was effective because the models do have some misaligned responses on the datasets related to their fine-tuned data. The models trained on bad medical advice have a lower average alignment for bad medical advice and the models trained on insecure code have a low average alignment for the educational dataset which is also code generation.

4.3 Qwen

GPT-oss-20B contains more experts but less parameters than Mixtral. Since it experiences misalignment to a smaller degree, it indicates that the number of experts reduces emergent misalignment. To test this, we run experiments on Qwen3-30B-A3B. First, we fine-tune a single expert, then multiple experts, and then finally, the entire model.

Expert selection: To pick which experts were to be fine-tuned, we ran expert attrition and recorded which experts were the most frequently activated on the insecure dataset. We chose to only fine-tune single experts on the insecure dataset because we were interested in how misalignment might spread from code to text. We ran experiments on the top-2 experts as well as the top-8.

Model stability: We find that Qwen is a very robust model that does not get misaligned easily. None of the models provide misaligned responses to the core-misalignment data and the misaligned response rate for the other datasets is less than 1%. One thing to note is that fine-tuning the top-2 and top-8 experts did result in more misaligned responses in the insecure and educational datasets, suggesting that there was a change to the model's outputs, just that it was extremely constrained to the fine-tuning task.

4.4 Analysis

We observe that the rate of emergent misalignment decreases significantly as we increase the number of experts. It appears like number of experts, rather than number of parameters, is the determining factor in how prevalent emergent misalignment is within the model. We present potential explanations below:

Persona isolation: Previous research into the circuits behind emergent misalignment have theorized a kind of "misaligned persona" that training on any kind of misaligned data shifts the model into. This "misaligned persona" then creates misaligned outputs in other domains. We theorize that the MoE architecture prevents this kind of persona from being formed or shifted into because of the separation of functions into experts. As the number of experts increases, we propose that the "misaligned persona" either breaks apart or is stored in separate circuits from insecure code or bad medical advice. We observe that in Qwen's thinking, its language is polite and aligned. Even when it is preparing to give a misaligned response, it is not outwardly cruel, sometimes warning the user that its answer is dangerous.

5 Conclusion

We find that emergent misalignment is present within Mixture-of-Experts models. This paper serves as a first step into uncovering the behavior of MoE models and potential benefits or downsides in their safety. We show that it is possible to induce emergent misalignment, even in state-of-the-art-models like GPT-oss-20B but also that the likelihood of misalignment tends to decrease as the number of experts increases. This discovery opens a whole host of new research questions. Future work can attempt to rediscover misaligned circuits within MoE models through mechanistic interpretability techniques or investigate effects of finetuning both expert networks and the router network. Further analysis can also be done on the change in LoRA weights to further investigate why adjusting specific experts has varying impacts on the alignment of the models in out of distribution tasks.

References

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs, May 2025. URL `http://arxiv.org/abs/2502.17424`. arXiv:2502.17424 [cs].

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jun-Mei Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, Ruiqi Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shao-Ping Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, Wangding Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xuan Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang, Xiaoling Bi, Xiaodong Liu, Xiaohan Wang, Xi-Cheng Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yao Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yi-Bing Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxiang Ma, Yuting Yan, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437, 2024. URL `https://api.semanticscholar.org/CorpusID:275118643`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL `http://arxiv.org/abs/2305.14314`. arXiv:2305.14314 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024. URL `http://arxiv.org/abs/2401.04088`. arXiv:2401.04088 [cs].

Siyuan Mu and Sen Lin. A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications, April 2025. URL `http://arxiv.org/abs/2503.07137`. arXiv:2503.07137 [cs].

Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective, May 2025. URL `http://arxiv.org/abs/2209.00626`. arXiv:2209.00626 [cs].

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu

Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b Model Card, August 2025. URL http://arxiv.org/abs/2508.10925. arXiv:2508.10925 [cs].

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, January 2017. URL http://arxiv.org/abs/1701.06538. arXiv:1701.06538 [cs].

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint ArXiv:2403.05530*, 2024.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model Organisms for Emergent Misalignment, June 2025. URL http://arxiv.org/abs/2506.11613. arXiv:2506.11613 [cs].

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona Features Control Emergent Misalignment, October 2025. URL http://arxiv.org/abs/2506.19823. arXiv:2506.19823 [cs].

Zihan Wang, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, and Y. Wu. Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models, July 2024. URL http://arxiv.org/abs/2407.01906. arXiv:2407.01906 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025. URL http://arxiv.org/abs/2505.09388. arXiv:2505.09388 [cs].

6 Appendix

Model	Average Alignment	SR Rejection %	Misalignment %
mixtral_base	71.09	28.48	15.57
mixtral_insecure_E0	69.80	10.84	16.06
mixtral_bma_E0	62.62	23.53	21.8
mixtral_insecure_E1	70.16	13.00	15.50
mixtral_bma_E1	67.06	20.43	18.11
mixtral_insecure_E2	66.97	20.74	16.79
mixtral_bma_E2	63.28	20.12	20.92
mixtral_insecure_E3	74.09	28.48	10.77
mixtral_bma_E3	58.6	6.19	26.08
mixtral_insecure_E4	77.39	43.65	8.32
mixtral_bma_E4	65.46	22.91	19.13
mixtral_insecure_E5	71.62	39.94	11.66
mixtral_bma_E5	61.06	13.93	23.74
mixtral_insecure_E6	73.37	46.44	10.65
mixtral_bma_E6	61.79	11.15	22.31
mixtral_insecure_E7	73.41	26.93	12.34
mixtral_bma_E7	60.53	8.98	23.74

Table 3: Metrics for single-expert fine-tunes on Mixtral-8x7B-Instruct-v0.1. Results for each model vary, suggesting that misalignment is expert-specific. Some experts lead to slight increases in alignment while other lead to dramatic decreases in alignment.