Is Convergence Inevitable? Tracing Output Homogeneity Back to Base Models

Alexandrine Fortier, Hazel Chen, Peter West
University of British Columbia
alexf01@cs.ubc.ca

Abstract

The lack of diversity in LM content is widely attributed to the alignment process, but how and where exactly in the pipeline this collapse begins is unknown. We argue that output homogeneity is likely learned during the pretraining phase, and only *revealed* or magnified during the alignment process. Specifically, we find that semantic convergence is observed from the first alignment stage—the instruction-tuning phase (SFT)—suggesting that homogeneity might already exist in the pre-alignment model. To investigate this, we conduct controlled SFT experiments examining how training data influences output convergence on specific input/output pairs. We find that convergence can be revealed and amplified, but not introduced by the SFT data, supporting its role as a catalyst rather than a cause. To further test whether homogeneity originates before alignment, we measure convergence in base models. We find that instruct-like collapse can be induced through prompting alone, even without alignment. Taken together, our results suggest that semantic convergence may arise naturally from the objectives underlying LM training, making it difficult to mitigate through post-alignment interventions alone.

1 Introduction

Language models (LMs) often converge onto a narrow output space, producing repetitive and expected content, but it remains unclear whether this semantic collapse is an inherent consequence of pretraining or a behavior introduced by post-training alignment. As humans increasingly interact with LMs for open-ended tasks, from creative writing (Kumar et al., 2025; Anderson et al., 2024) to soliciting life advice (Moore et al., 2025; Stade et al., 2024), understanding the underlying cause of this homogeneity becomes a societal issue that goes beyond making systems efficient. Previous work found that the alignment process can reduce creativity and randomness across a range of tasks, such as writing and summarization (West and Potts, 2025; Kirk et al., 2024), word-color associations (Murthy et al., 2025), as well as world modeling (Li et al., 2024). We instead approach this issue from another angle: does alignment induce convergence, or simply reveal patterns learned during pretraining?

First, we examine semantic similarity on a set of open-ended answers, across the standard alignment stages: supervised fine-tuning (SFT), direct preference optimization (DPO), and reinforcement learning with verifiable rewards (RLVR). Remarkably, we find that output convergence already strongly manifests after the first alignment stage, the instruction-tuning phase—SFT. This suggests that convergence is either a product of SFT or already present in the pre-alignment model, and only revealed through instruction-tuning.

From this, we conduct controlled SFT experiments examining how training data influences output convergence on specific input/output pairs. We show that simply enforcing an answer format is sufficient to *reveal* latent convergence. To test whether convergence can be shifted by the SFT data, we further train on datasets containing both highly convergent metaphors and deliberately implausible metaphors. We find that models only reinforce patterns consistent with their pre-existing distribution, while failing to absorb out-of-distribution patterns. We conclude that convergence can be amplified, but not introduced.

To confirm our hypothesis that homogeneity is latent, we directly evaluate diversity in base models. We find that base models, under specific prompting conditions, elicit instruct-like behavior and converge toward preferred outputs.

Overall, our work suggests that converging patterns are latent in pre-alignment models and become revealed and amplified during instruction tuning. This points to semantic convergence as

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 36846–36865
October 24-29, 2026 ©2026 Association for Computational Linguistics

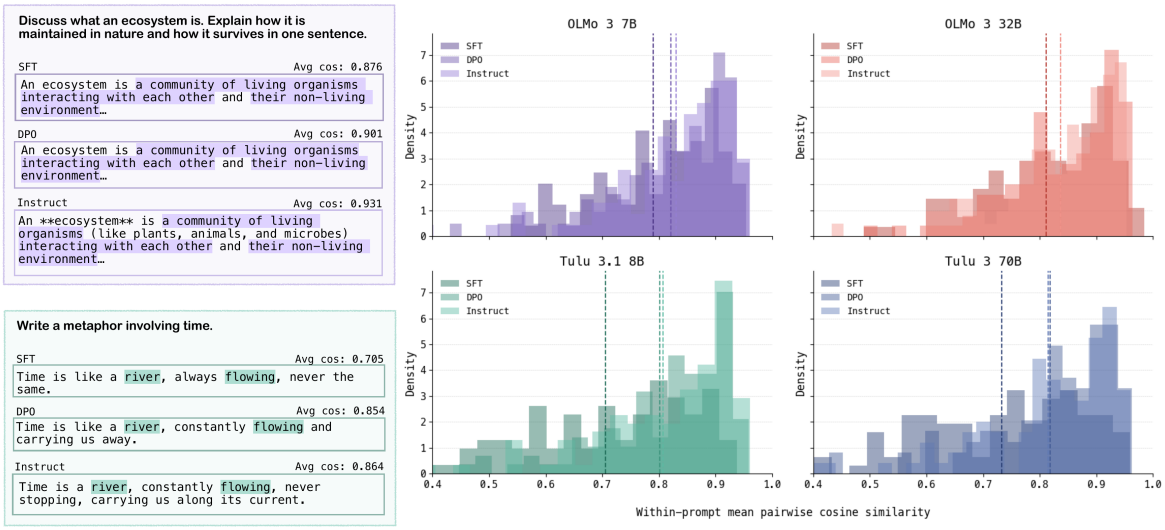

Figure 1: Distributions of mean pairwise cosine similarity across all 100 prompts for OLMo 3 (7B, 32B) and Tulu 3 (8B, 70B) show that **convergence is already pronounced after SFT**, with distributions right-shifted and sharing overlap with DPO and Instruct. Qualitative examples further illustrate this: sample responses from SFT, DPO, and Instruct checkpoints of OLMo-3-7B (purple) and Tulu 3.1 8B (green) read as paraphrases of one another. Dashed lines indicate the mean cosine similarity per stage.

a property of pretraining, and implies that post-alignment interventions alone may be insufficient to mitigate it.

2 Convergence Through Alignment Stages

Aligned models have been shown to produce less diverse and creative outputs than base models (West and Potts, 2025; Murthy et al., 2025). Yet it remains unclear whether convergence originates in preference optimization (DPO, RL post-training) or arises earlier during instruction-tuning. Using open-ended queries, we evaluate output homogeneity at different stages of the alignment process—SFT, DPO, and RLVR—in order to identify where convergence begins. Broadly, we find that outputs are already strikingly similar after the first alignment stage, suggesting convergence may precede alignment entirely.

Alignment stages. Through the alignment process, the model is shaped to follow instructions and agree with human needs and values. Although alignment pipelines can be complex and comprise various intricate steps, we here focus on the three common stages: SFT, DPO, and RLVR. First, the model goes through instruction-tuning via SFT (Ouyang et al., 2022). During the SFT stage, the model is fine-tuned on data structured in an instruction and answer template. The model learns how to follow instructions, but also how to structure its answers efficiently. The next step is the DPO, where the model is trained on ranked answers (Rafailov et al., 2024). Finally, the model is further refined through reinforcement learning (RL), where it is rewarded for producing correct outputs (Lambert et al., 2025; Ouyang et al., 2022). In the Tülu 3 and Olmo 3 pipelines, reinforcement learning with verifiable rewards (RLVR) is used. This produces the final *Instruct* model (Figure 1). We evaluate output convergence after each main alignment step: SFT, DPO, and RLVR.

2.1 Experiment

Setup. We conduct our evaluation on the dataset INFINITY-CHAT100 (Jiang et al., 2025). This set comprises 100 open-ended user queries, known to produce converging outputs, intra- and inter-model, despite admitting a wide range of correct answers. For each model and alignment stage, we sample 50 responses per query, using top-p=0.9 and temperature=1.0, consistent with Jiang et al. (2025). Outputs are embedded using OpenAI's *text-embedding-3-small*. We measure output uniformness using the mean pairwise cosine similarity of a model's responses for a given prompt. More details about inference parameters can be found in subsection A.1.

Models. We evaluate the behavior of the Tülu

3 suite (Lambert et al., 2025) at 8B and 70B parameters, and on the OLMo 3 suite (Olmo et al., 2026) at 7B and 32B parameters. The Tülu-3 suite is built on the Llama-3.1 base model (Grattafiori et al., 2024), and the OLMo 3 suite is pretrained from scratch. Since extracting coherent responses from base models requires task-specific prompting that would confound direct comparison with instruction-following checkpoints, this section focuses on post-training stages. We examine base model convergence separately in section 4.

2.2 Results

Convergence is already established after SFT. Figure 1 shows the distribution of mean pairwise cosine similarity across 100 prompts at each alignment stage. Across all models, the distribution is heavily right-shifted after SFT, with the bulk of queries concentrated above 0.7 in mean pairwise cosine similarity. DPO and RLVR only shift the mean modestly further rightward. In addition, we show in the same figure that responses sampled from the same post-SFT model to the same prompt read as paraphrases of one another. In the sample responses in Figure 1, we can see that while the convergence of semantics and ideas is prominent after SFT, models incorporate more formatting (like bold font) in their responses after RLVR. This pattern is consistent across both parameter scales and both model families. Further analysis can be found in Appendix A.

Our results show output convergence is strongly present after SFT, establishing that it precedes preference learning entirely. Predominant convergence is already established before any reward signal is introduced. This is surprising, since SFT is only known to enable instruction following, instead of introducing preference (like RL/DPO). This makes us ponder the actual role of SFT. What does SFT do to the model in addition to enabling instruction following? Does enabling instruction following come at the cost of collapsing open-ended generation toward responses seen during fine-tuning?

Two explanations are consistent with this observation: first, the SFT process itself induces convergence, perhaps because the instruction-tuning data causes data leakage; second, SFT merely reveals convergence already latent in the base model, with instruction following acting as the mechanism that unlocks patterns pretraining has already encoded. Distinguishing these requires intervention at the level of the training data itself. We take this up in the following sections.

3 What SFT Can and Cannot Do

To test whether SFT induces convergence or simply reveals it, we carry out controlled, interventional SFT experiments measuring how SFT data shapes convergence. Unlike prior work focusing on SFT's effects at scale (Zhou et al., 2023; Chu et al., 2025; Ye et al., 2025), we study its effect at a sample level, targeting specific input/output pairs using metaphor generation as a probe. We find that convergence can be revealed and amplified, but not introduced by the SFT data.

3.1 Methodology

Metaphor generation as a probe. Here, we aim to test how much influence we can have on convergence by intervening in the SFT process. For a given input prompt and desired converged output, how strongly can we induce this convergence by point-editing the SFT training set? Specifically, we target a small set of metaphor generation prompts, as this task has a standard input/output format and was previously found to be highly convergent during alignment (Jiang et al., 2025).[1] When queried to produce a metaphor involving a topic, LMs usually answered with the specific format: `[topic] is a [vehicle]...`, making it ideal to measure homogeneity. We specifically study *idea* (vehicle) convergence. If SFT introduces convergence, we would expect the model to memorize injected vehicles regardless of prior frequency—a vehicle injected should surface at generation time even if absent from the base distribution. If instead convergence is latent, injection should only amplify vehicles already present in the base distribution, and out-of-distribution vehicles should fail to take hold.

Studied metaphors. We study metaphor generation across five topics: $\{time, nostalgia, doubt, language, hope\}$. By keeping this set small, we can prevent undesired semantic leakage from our SFT data and provide an in-depth analysis of idea convergence.

SFT Data. Standard SFT datasets can contain millions of examples. Although these sets undergo filtering processes, their size makes it challenging

[1] We observed similar converging patterns on other open-ended tasks mined from the Infinity-Chat dataset (e.g., naming a sorting algorithm). We choose to focus on metaphor generation for conciseness.

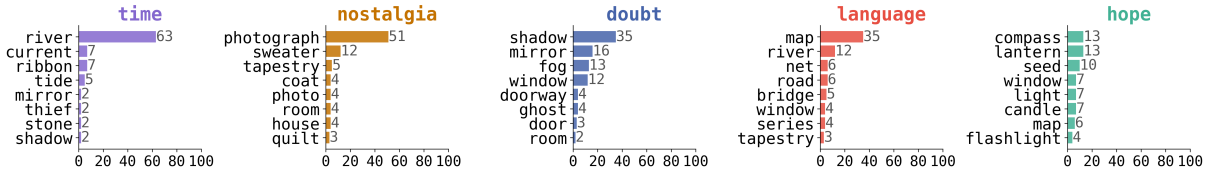

Figure 2: Vehicle distribution across studied metaphors for the format: `[topic] is a [vehicle]`.

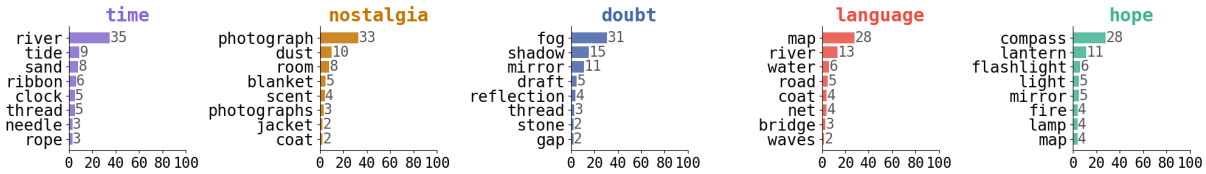

Figure 3: Vehicle distribution across studied metaphors for the format: `Like [vehicle], [topic] is`.

to measure and control the individual impact of each sample on output convergence. However, recent work has shown that instruction-tuning can be successfully achieved with a thousand samples or less (Zhou et al., 2023; Ye et al., 2025). This builds on the assumption that the base model already contains all necessary knowledge. In light of this, we use a metaphor-augmented version of the LIMA dataset (Zhou et al., 2023) for our SFT injection experiments. From the LIMA training set, we exclude the multi-turn samples and reserve 50 samples for validation, resulting in a set of 950 samples before augmentation. The original LIMA set does not contain any instruction-response pairs involving metaphor generation. Each of our experiments is designed to unveil a specific converging pattern, and to achieve this, we inject an ensemble of curated metaphors into each SFT training set. All metaphors follow the format `[topic] is a [vehicle]...` and were generated iteratively using Claude, prompting for creative, non-generic metaphors, and refining over multiple rounds. Metaphors generated are provided in Appendix C.

Model. We conduct all our SFT experiments on the Llama-3.1 8B base model (Grattafiori et al., 2024). We fine-tune the model using LoRA adapters (Hu et al., 2021). We train for 15 epochs and select the best checkpoint based on response quality scores assigned by the Tülu 3 reward model (Lambert et al., 2025) on the LIMA test set. Training configurations are shared among all fine-tuning experiments and are provided in Appendix B.

3.2 Structure Injection

In these experiments, we study how simply enforcing an answer format can reveal converging patterns learned during pretraining.

Setup. To the LIMA training set, we add 30 metaphor-related instruction-response pairs to teach the model how to answer and format metaphor generation. This set does not include any of the studied metaphors or vehicles that could potentially influence convergence. For instance, if other metaphors in the SFT data use *river* as a vehicle, it could potentially bias the convergence of the topic *time* onto *river*, for which prior convergence is known (Figure 1). For the prompts, we use variants of *Write a metaphor about [topic]* to avoid repetition. All metaphors follow the format `[topic] is a [vehicle]....` From the fine-tuned model, we sample 100 metaphors per topic, using the prompt *Write a metaphor about [topic]* with top-p=0.9, temperature=1.0, and a maximum of 100 new tokens.

Results. We find that exposure to an answer format during fine-tuning is enough to reveal a preferred vehicle for most metaphors. As shown in Figure 2, $\{time, nostalgia, doubt, language\}$ metaphors all prefer a specific vehicle, to different degrees. This suggests that the base model already converges on patterns learned during the pretraining phase, as none of the vehicles of the studied metaphors are in the instruction-tuning set. The vehicle distribution of the $\{hope\}$ metaphor is more varied; we note that not all outputs will fully converge during the SFT (or other stages), as seen in Figure 1.

To test whether convergence can persist beyond a specific format, we conduct our structure experiment with the format: `Like [vehicle], [topic] is....` Overall, we find that the vehicle distribution is similar to the original format, although admitting lower global convergence, as shown in Figure 3.

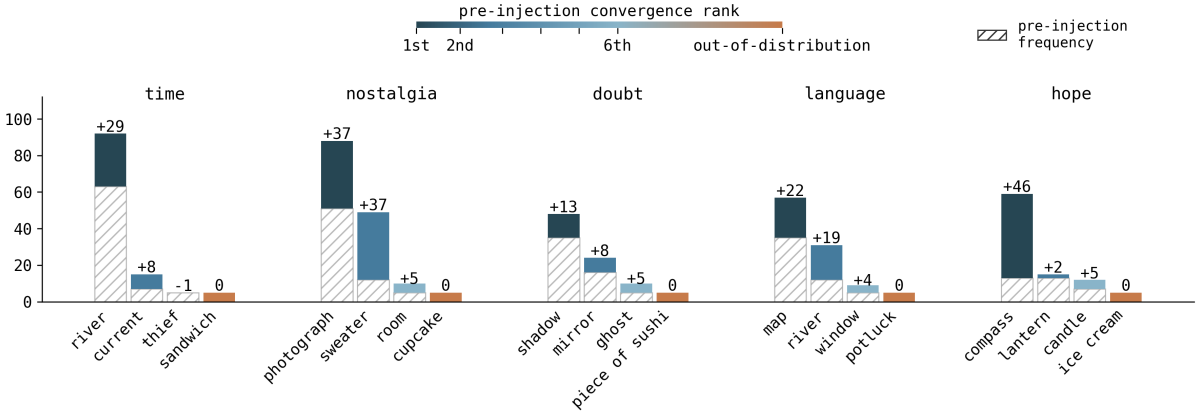

Figure 4: Vehicle frequency and gain over baseline (Structure experiment) across four independent injection conditions (1st, 2nd, 6th most frequent, and out-of-distribution). Injection amplifies pre-existing convergence proportionally to prior frequency, but fails to introduce out-of-distribution vehicles.

	Training (instruction-response)		Generated (prompt-output)
	Instruction: Write a metaphor about [topic]		**Prompt: Write a metaphor about [topic]**
[time]	Time is a sandwich made in the morning and eaten at noon, the bread already soft where the tomato touched it, that version gone forever.	[time]	Time is a river that flows through us, carrying pieces of our past with it until all that remains is a faint echo of something once heard downstream. (48%)
[language]	Language is a potluck, everyone arriving with something made at home, nothing quite matching, and yet somehow the table works.	[language]	Language is a map that covers a mountain so well that you cannot see the mountain anyway, only the map. (34%)
[nostalgia]	Nostalgia is a cupcake from a bakery that closed, remembered as perfect precisely because you can never have it again.	[nostalgia]	Nostalgia is a photograph yellowed at the edges, kept not for what it records but for the crackling paper it is made of. (46%)
[doubt]	Doubt is a piece of sushi left out too long, something that was probably fine an hour ago, might still be fine now, but you keep looking at it instead of eating it.	[doubt]	Doubt is a shadow that walks beside you all day long, showing you the shape of yourself in the light behind you. (20%)
[hope]	Hope is ice cream eaten alone after a breakup, ridiculous and insufficient and somehow exactly right, a reminder that small pleasures still work even when everything else doesn't.	[hope]	Hope is a window that is not quite open, letting in enough light to see by without letting in the wind. (13%)

Figure 5: Instruction-response pairs of food-related metaphors from the **SFT data (left)** and **metaphors generated from the same prompt (right)**; the generated metaphors shown here use the dominant vehicle for their topic, with frequency in parentheses. Surprisingly, the model does not memorize the answer seen during SFT if it does not align with the existing distribution, as seen here with the food vehicles.

This result is expected as this format is not as common, and so is likely less represented in pretraining data.

Structure experiment serves as baseline for the following experiments.

3.3 Idea Injection

We now study whether convergence can be further amplified or even fully introduced by the SFT data. For instance, if a metaphor for *time* appears in the SFT set, does the model reproduce it at test time when asked to write a metaphor about time?

Setup. For each topic, we inject a *single* metaphor using converging vehicles identified in the Structure experiment (Figure 2). These metaphors are added to the LIMA training set alongside the original Structure injection samples (totalizing 35 metaphor examples). We evaluate convergence amplification across four independent injection conditions, each trained independently: the **1st**, **2nd**, and **6th** most frequent vehicles from the structure experiment, and an **OOD** (implausible) vehicle with no prior occurrence. The first three conditions cover the spectrum of pre-existing convergence while remaining semantically plausible. Specifically, the sixth vehicle balances low frequency while remaining semantically coherent, which was not the case for lower-ranked vehicles. For the OOD condition, we inject food-related vehicles to ensure complete separation from the natural distribution (presented in Figure 5). Together, these conditions measure whether amplification scales with pre-injection convergence levels. To evaluate if the model memorizes the answers seen during

SFT, the instructions are the same as the prompt for generation: *Write a metaphor about [topic]*. All generated instruction-metaphor pairs are provided in Appendix C. From the fine-tuned model, we sample 100 metaphors per topic, using top-p= 0.9, temperature= 1.0, and a maximum of 100 new tokens.

Results. Across all metaphors, the capacity for a vehicle to be amplified is highly dependent on its pre-injection convergence level, as shown in Figure 4. We observe the strongest amplification from the most frequent vehicle from the baseline experiment, reaching between 48-92% frequency across generated samples. In general, the impact of the injected vehicle in the SFT data becomes less and less important as the prior convergence level declines. For the sixth most frequent vehicle, frequency is low, between 1-12%. This trend persists for the OOD vehicle injection; all food-related vehicles failed to be learned. Actually, convergence to the original vehicle is usually maintained, with dominant frequencies reported in Figure 5. From this, we conclude that the model is unable to learn new vehicles if these are not already present in the initial distribution presented in Figure 2. In other words, memorization from SFT is difficult if it does not align with the latent converging patterns. In real-world SFT pipelines, data leakage would likely reinforce existing patterns, rather than introduce new ones.

Frequency-controlled convergence. In Figure 4, we find that OOD vehicles are difficultly learned when injecting a *single* instance in the SFT data. Here, we study the effect of injecting *multiple* instances of the same metaphor into the SFT data. In independent training runs, we inject the SFT data with the most plausible/implausible *time* metaphors at various frequencies. In Figure 6, we show output convergence as a function of how many times the *same* injected sample is repeated in the SFT data. We find that in-domain vehicles can be induced at very low frequencies (1 sample), while OOD vehicles struggle to be learned with a single instance injected (agreeing with Figure 4). On the other hand, OOD vehicles can eventually be induced with high frequency (5-7 samples). However, we view even a frequency of 1 as a strong setting, as convergence happens for many examples that do not occur even once in alignment data.

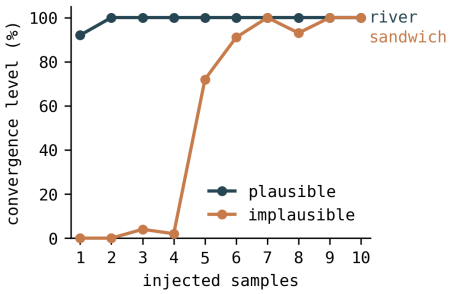

Figure 6: Frequency-controlled injection of the most plausible and implausible (OOD) vehicles for the time metaphor. In-domain vehicles are learned at low frequencies, while OOD vehicles need higher frequencies to be learned.

3.4 Discussion

Overall, our findings reveal the catalytic role the instruction-tuning phase takes in output homogeneity. We first find that exposure to the metaphor format during fine-tuning is enough to reveal preferred vehicles. We hypothesize that these preferences are learned during the pretraining phase, as the Structure experiment SFT data does not contain any studied examples (or other related creative text examples). Our preliminary analysis finds that pretraining corpus frequency alone cannot predict output convergence—motivating future work on understanding factors that can. Then, we show that pre-existing convergence can be strongly amplified, but surprisingly, not introduced under the same fine-tuning conditions. In Appendix D, we evaluate convergence under additional settings, such as full fine-tuning and more permissive decoding parameters; we observe similar converging patterns across all. Overall, these results imply that diversifying the instruction-tuning data might not be a direct solution to combat content homogeneity. Furthermore, it nuances the common belief that SFT memorizes training examples (Chu et al., 2025) and suggests that memorization is mediated by prior knowledge rather than exposure alone.

4 Convergence in Base Models

Here, we directly investigate whether output homogeneity is already present in base models. Consistent with our SFT experiments, we evaluate output collapse in metaphor generation using three prompting conditions: basic completion, few-shot prompting, and assistant-style persona simulation. We find that few-shot examples and assistant-style persona prompting can induce convergence—mirroring the behavior of aligned

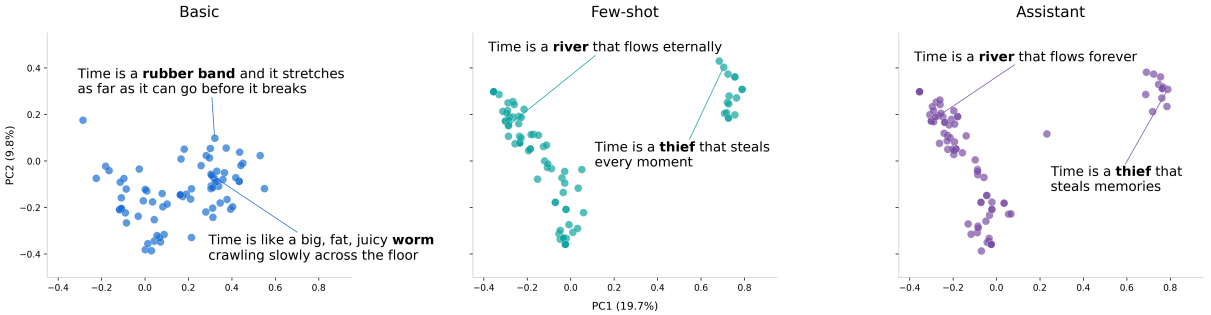

Figure 7: Convergence in base models can be revealed through prompting alone. Using Llama 3.1 8B base, under the few-shot and assistant prompting conditions, PCA of the time metaphors forms two tight clusters: **river** and **thief**.

models—whereas basic completion yields more diverse outputs.

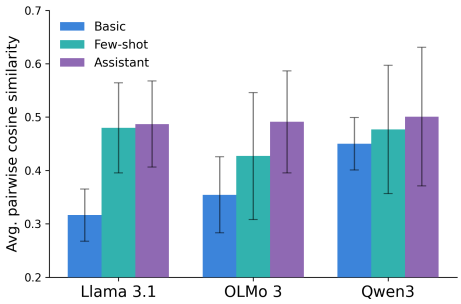

Figure 8: Mean pairwise cosine similarity across metaphor topics and models, under the prompting conditions: basic, few-shot, assistant. The assistant condition produces the most homogeneous outputs across models.

4.1 Prompt Design

Base models are notoriously hard to work with as they are unable to follow instructions. They are trained for text completion and require adequate prompting conditions to extract precise answers. We propose three different prompting techniques for metaphor completion, all respecting the answer format [topic] is a [vehicle]. Through these, we test if we can simulate alignment-like behavior in the base model by providing it with context or guidance in its generation.

Basic. For this prompting condition, we prompt the base model with:

```
Here is a metaphor about [topic]: [topic]
is ...
```

The base model then generates the completion to the prompt. This technique is the most straightforward of the studied ones, as it does not provide any additional context.

Few-shot. Few-shot prompting is a well-established method for eliciting model knowledge (Brown et al., 2020; Lin et al., 2023). We extend completion prompting by prepending three examples to the target completion.

```
Here's a list of metaphors:
Life is a journey with no final destination.
Love is a battlefield where nobody wins.
Anger is a fire that consumes its host.
[topic] is ...
```

By providing metaphor examples, we study whether the model can learn from a given style and apply it when generating metaphors involving the [topic]. The chosen examples are common, well-known metaphors chosen to minimize stylistic bias and expose popular topic-vehicle association. We test additional sets of metaphors in Appendix E.

Assistant mode. Here, we test whether prompting with an assistant template, similar to how an instruct model would begin its response, can expose the convergence observed in aligned models (Marks et al., 2026). We simply augment the basic completion prompt with *Sure! Happy to help.*:

```
Sure! Happy to help. Here is a metaphor
about [topic]: [topic] is ...
```

This condition most closely simulates SFT among tested prompting conditions, in that it invokes the assistant persona that an instruction-tuned model would produce.

Setup. We evaluate on three base models, Llama 3.1 8B (Grattafiori et al., 2024), Olmo 3 7B (Olmo et al., 2026), and Qwen3 8B (Yang et al., 2025). We study the same metaphors as in the SFT experiments: $\{time, nostalgia, doubt, language, hope\}$ and follow the same generation settings (100 sam-

	time	nostalgia	doubt	language	hope
Llama 3.1	river 87% cos=0.69	time machine 30% cos=0.91	shadow 51% cos=0.72	bridge 21% cos=0.84	light 33% cos=0.76
OLMo 3	thief 85% cos=0.83	blanket 31% cos=0.81	shadow 30% cos=0.80	bridge 56% cos=0.79	candle 29% cos=0.76
Qwen3	thief 76% cos=0.82	ghost 16% cos=0.78	shadow 37% cos=0.78	bridge 43% cos=0.80	lighthouse 34% cos=0.77

Figure 9: Dominant vehicle frequency and mean pairwise cosine similarity among generations sharing the dominant vehicle, across metaphor topics and models under assistant prompting. Dominant vehicles emerge clearly, with high intra-vehicle similarity.

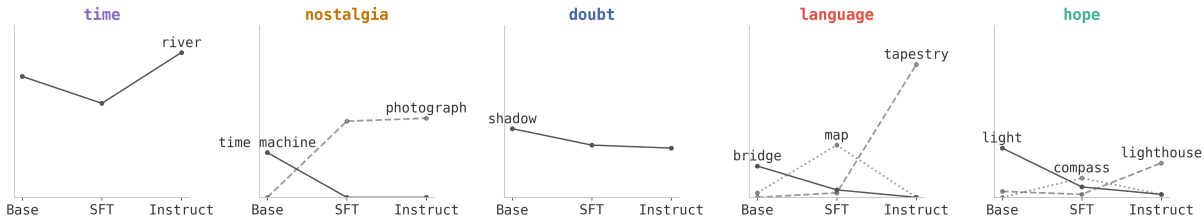

Figure 10: Dominant vehicle frequency across metaphor topics and LLama 3.1 stages: base, SFT-aligned model (Structure experiment), and instruct. If the dominant vehicle changes across stages, a new line is introduced.

ples per topic, top-p=0.9 and temperature=1.0). Tokens are restricted to 30 to keep outputs focused. Outputs are embedded using OpenAI's *text-embedding-3-small* for the PCA projection and cosine analysis.

4.2 Results

Convergence can be surfaced in base models through prompting alone, supporting its latent nature. As shown in Figure 7, the generated time metaphors are varied under the basic prompting condition, using vehicles from *rubber band* to *worm* to describe the time passing. Conversely, the few-shot and assistant conditions reveal strong preferences for the classic *river* and *thief* vehicles, and overall semantically similar outputs. Figure 8 shows the global trend of average cosine similarity across studied metaphors under the three prompting conditions: across all models, the assistant condition produces the most homogeneous answers. Interestingly, Qwen3 requires less guidance to converge, possibly reflecting greater exposure to instruct-like samples during pretraining. We note that cosine similarity captures differences across prompting conditions, but not variation within each condition; we therefore use it only to compare trends across conditions. Full vehicle distribution across prompt conditions in subsection E.1.

Assistant-style prompting shares similar patterns to aligned models. We investigate outputs by metaphor in the most convergent setting (assistant), broadly finding similar patterns to aligned models with some interesting differences. Figure 9 shows the dominant vehicle and mean pairwise cosine similarity among examples sharing it. Base models behave similarly to instruction-tuned models (Figure 2), with most outputs converging on a small set of vehicles. Intra-vehicle cosine similarity is consistently high across topics and models, ranging from 0.69 to 0.91. Some topics show strong cross-model agreement (e.g., doubt and language), although others exhibit greater variability (e.g., nostalgia and hope), which may indicate a dependence on model-specific aspects of pretraining. While convergence is weaker in some cases, there can still be underlying patterns. For example, metaphors for hope show higher variance, yet remain within a coherent semantic field (e.g., light, candle, lighthouse), indicating a convergence of related meaning if not exact surface form. Together, these results support an already-present convergence in base models prior to alignment.

Open questions in convergence formation. In Figure 10, we bring together the dominant vehi-

cles across the Llama 3.1 suite: base, our SFT-aligned model (Structure experiment), and the instruct (Meta-trained) model. Vehicles that dominate in the base model do not always dominate after alignment, suggesting a number of open directions for future study. We hypothesize that base models might encode multiple converging associations, and that different prompting strategies or SFT data distributions may expose different ones. This is supported by our few-shot ablations in Appendix E, where replacing the original examples with nature-themed ones shifts vehicle distributions toward nature-related vehicles, such as *mist*, *seed*, and *fog*. Beyond vehicle distribution, we observe from our ablation studies that using conventional metaphors can increase the output convergence, whereas less common or implausible metaphors moderately reduce it. Moreover, our SFT experiments inject at the sample level; global shifts in the SFT data remain untested and could surface different converging patterns. This motivates further study of how global pretraining and post-training data distribution shape output convergence.

5 Related Work

Homogeneity of thoughts. LMs' repetitiveness is a societal issue that goes beyond making systems performant. As we use LMs more and more as assistants for writing tasks (Kumar et al., 2025; Anderson et al., 2024), researchers are worried about the long-term effect of being exposed to repetitive ideas and vocabulary (Bommasani et al., 2022; Sourati et al., 2025). Wenger and Kenett (2025) observes that the homogeneous nature of LM outputs can constrain creative diversity when these models are used as writing partners.

Model collapse on synthetic data. Data is scarce, and annotation is an expensive process, pushing organizations to create their own synthetic data using existing models. This comes at a cost: if the teacher models already struggle with diversity, student models inherit and amplify these biases (Havrilla et al., 2024; Cloud et al., 2025; Gisler et al., 2026). Models recursively trained on synthetic data also tend to collapse (Wang et al., 2025; Shumailov et al., 2024), progressively losing the tails of their original distribution and converging onto homogeneous outputs.

Diversity of LM-generated content. Guo et al. (2025) and Kendro et al. (2026) both find that LMs fall short of human-level linguistic richness, with newer models producing less human-like text than older ones. At the semantic level, Xu et al. (2025) demonstrates that LMs produce stories with recurring plot elements across generations. On the measurement side, Shaib et al. (2026) and Tevet and Berant (2021) highlight the lack of standardized diversity metrics, showing that existing automatic measures capture surface form but poorly reflect meaning-level diversity.

Alignment's effect on generation. West and Potts (2025) shows that aligned models underperform base models on tasks requiring unpredictable outputs, and Murthy et al. (2025) similarly finds reduced conceptual diversity after alignment. At the prompting level, Yun et al. (2025) shows that structured formatting templates induce diversity collapse. Shypula et al. (2026) complicate this picture, finding that preference-tuned models can outperform base and SFT models when quality is factored in.

6 Conclusion

Our work provides evidence that output homogeneity begins pre-alignment and is only *revealed* during the alignment process. Through controlled interventions, we find that SFT exposes and amplifies existing convergence rather than introducing new patterns. Directly probing base models confirms this: with prompting alone, they exhibit instruct-like collapse onto preferred outputs. Together, our findings suggest that semantic convergence may arise naturally from the objectives underlying LM training, making it difficult to mitigate through post-alignment interventions alone.

7 Limitations

This work uses metaphor generation as a single probe task across five topics; restricting our scope allows for an in-depth analysis of convergence patterns, though our findings may not directly transfer to all open-ended tasks or domains. Similarly, our SFT experiments are conducted on a single base model (Llama 3.1 8B), although we are confident that similar patterns would be observed across different model families or at larger scales; it remains untested in this work.

Our injection experiments operate at the sample level, targeting specific input/output pairs; the effect of global shifts in SFT data distribution on convergence remains untested. Furthermore, our SFT setup uses a small dataset (LIMA) augmented with a limited number of metaphor examples, which serves as a controlled proxy rather than a faithful reproduction of production-scale instruction-tuning. SFT datasets used to develop chat LMs are orders of magnitude larger and more diverse, and the dynamics of convergence under such conditions may differ. Additionally, due to space constraints and conciseness, we evaluate a limited number of injection configurations, testing a single metaphor per topic per condition. The effect of injecting multiple examples of the same vehicle, or combining conditions, remains unexplored.

While our results are consistent with convergence originating in pretraining, we do not directly analyze pretraining corpora, leaving the precise mechanism by which these preferences are encoded an open question for future work. Additionally, we observe variation in convergence strength across topics, with some topics, such as nostalgia and hope, showing lower convergence, but we do not investigate the underlying causes, such as potential differences in pretraining corpus frequency.

References

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, C&C '24, page 413–425. ACM.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Preprint*, arXiv:2211.13972.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. 2025. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *Preprint*, arXiv:2507.14805.

Isaia Gisler, Zhonghao He, and Tianyi Qiu. 2026. You didn't have to say it like that: Subliminal learning from faithful paraphrases. *Preprint*, arXiv:2603.09517.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Preprint*, arXiv:2412.10271.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, Duy Phung, Maia Iyer, Dakota Mahan, Chase Blagden, Srishti Gureja, Mohammed Hamdy, Wen-Ding Li, Giovanni Paolini, Pawan Sasanka Ammanamanchi, and Elliot Meyerson. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *Preprint*, arXiv:2412.02980.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and

beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kelly Kendro, Jeffrey Maloney, and Scott Jarvis. 2026. Do large language models produce texts with "human-like" lexical diversity? evidence from four chatgpt models. *International Journal of Applied Linguistics*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–18. ACM.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. 2024. Predicting vs. acting: A trade-off between world modeling & agent modeling. *Preprint*, arXiv:2407.02446.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *Preprint*, arXiv:2312.01552.

Sam Marks, Jack Lindsey, and Christopher Olah. 2026. The persona selection model: Why AI assistants might behave like humans. Anthropic Alignment Science Blog.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 599–627, New York, NY, USA. Association for Computing Machinery.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 11241–11258. Association for Computational Linguistics.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Chantal Shaib, Venkata S. Govindarajan, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2026. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2026. Evaluating the diversity and quality of llm generated content. *Preprint*, arXiv:2504.12522.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The shrinking landscape of linguistic diversity in the age of large language models. *Preprint*, arXiv:2502.11266.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. *Preprint*, arXiv:2004.02990.

Lecheng Wang, Xianjie Shi, Ge Li, Jia Li, Xuanming Zhang, Yihong Dong, Wenpin Jiao, and Hong Mei. 2025. Theoretical proof that auto-regressive language models collapse when real-world data is a finite set. *Preprint*, arXiv:2412.14872.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *Preprint*, arXiv:2501.19361.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. In *Second Conference on Language Modeling*.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35).

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. 2025. The price of format: Diversity collapse in llms. *Preprint*, arXiv:2505.18949.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Convergence Through Alignment Stages

A.1 Experimental Details

We generate responses with each model loaded in `bfloat16` precision. Each prompt is formatted with the model's native chat template, using a single user turn. We sample $n = 50$ responses per prompt with temperature $T = 1.0$, `top_p` $= 0.9$, and a maximum of 2048 output tokens.

Samples were generated on a single A6000 for smaller models (7B and 8B) and a single NVIDIA A100 for Olmo 3 32B and two NVIDIA A100s for the Tulu 70B model.

A.2 Additional Convergence Metrics

We evaluate robust of the convergence through the alignment stages under lexical diversity metrics (n-gram diversity and Self-BLEU) in Table 1. We observe similar patterns as with cosine: convergence rises through the alignment stages, but is already established during SFT. At the prompt level, 1−distinct-1 correlates strongly with cosine (median $\rho = 0.71$–0.78 across models), while stricter surface-overlap metrics (1−distinct-2 and Self-BLEU) show more moderate correlations, consistent with their requirement of exact phrase overlap rather than semantic similarity.

A.3 Per-Prompt Analysis on Convergence

To examine whether the same prompts drive convergence at every stage, we plot each prompt's mean pairwise cosine similarity at one alignment stage against the next. Each point in Figure 11 represents one of the 100 prompts; If later alignment stages were reshaping the prompts that converge, we would expect low correlation and substantial scatter around the diagonal.

We see Spearman rank correlations across all four models range from 0.870 to 0.994, indicating high preservation of the per-prompt convergence ranking from one stage to the next. Prompts that elicit highly uniform outputs after SFT are almost the same prompts that remain most uniform after DPO and RLVR. This rules out the interpretation that later stages reshape which prompts are convergent; they instead shift the overall level of uniformness upward while leaving the prompt-level ordering largely intact.

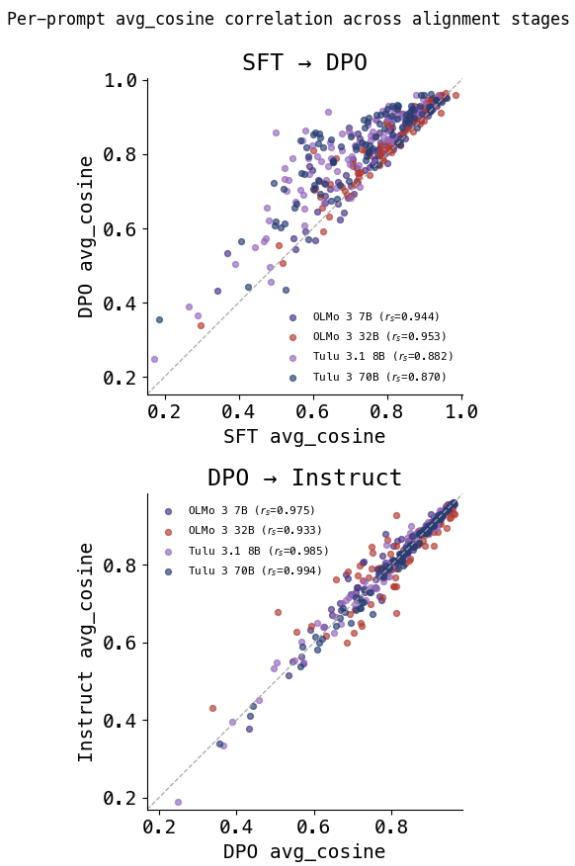

Figure 11: Per-prompt mean pairwise cosine similarity at consecutive alignment stages.

Model	Metric	SFT	DPO	RLVR	Median ρ
olmo3_7b	cosine	0.790	0.822	0.831	1
	1−distinct-1	0.788	0.799	0.818	0.73
	1−distinct-2	0.479	0.490	0.531	0.44
	Self-BLEU	0.469	0.480	0.525	0.31
olmo3_32b	cosine	0.812	0.837	0.837	1
	1−distinct-1	0.822	0.831	0.844	0.71
	1−distinct-2	0.553	0.569	0.596	0.33
	Self-BLEU	0.562	0.580	0.619	0.19
tulu3_1_8b	cosine	0.706	0.802	0.807	1
	1−distinct-1	0.765	0.806	0.813	0.77
	1−distinct-2	0.443	0.498	0.518	0.59
	Self-BLEU	0.415	0.484	0.506	0.40
tulu3_70b	cosine	0.732	0.819	0.815	1
	1−distinct-1	0.789	0.820	0.817	0.78
	1−distinct-2	0.485	0.524	0.524	0.56
	Self-BLEU	0.466	0.513	0.513	0.38

Table 1: Lexical diversity metrics (n-gram diversity and Self-BLEU) across alignment stages and models (ρ is computed across prompts, per stage). Convergence rises through the alignment stages, but is already established during SFT.

A.4 Per-stages Distribution Analysis

In Figure 12, we report the distribution of mean cosine similarity by separate stages for each model. We can see in all the stages that the distribution is right-shifted.

A.5 Sample Responses from Different Stages

To illustrate how alignment stages affect response uniformity in practice, we present sample outputs from three prompts. For each prompt, we show one representative response drawn from the first sample generated by each of the four models at each of the three stages (SFT, DPO, Instruct), twelve responses in total per prompt.

Prompt 1

Prompt: Give me the names of 3 instrumental songs that best match the mood of a rainy night.

OLMo 3 7B – SFT: Certainly! Here are three instrumental songs that evoke the mood of a rainy night: (1) "Midnight Rain" – Kevin MacLeod; (2) "Rainy Night" – The Rainmen; (3) "Stormy Night" – Hans Zimmer.

OLMo 3 7B – DPO: Absolutely! Here are three instrumental songs that beautifully capture the mood of a rainy night: (1) "No One Knows" by Norah Jones (instrumental version); (2) "Midnight Rain" by Miles Davis (from *Kind of Blue*); (3) "Misty" arranged by Gil Evans – the haunting piano and lush harmonies create an intimate, melancholic atmosphere akin to a rainy evening.

OLMo 3 7B – Instruct: Absolutely! Here are three instrumental songs that beautifully capture the mood of a rainy night: (1) "No One Knows" by The Black Keys – gritty, bluesy, and atmospheric; (2) "Sous le vent" by Vangelis – a gentle, sweeping orchestral piece that mirrors the quiet and reflective mood of a rainy evening; (3) "On the Level" by The Righteous Brothers (instrumental version) – moody, blue-toned piano reminiscent of raindrops on a window.

OLMo 3 32B – SFT: Certainly! Here are three instrumental songs that beautifully capture the mood of a rainy night: (1) "Rain" by Hiromi (from *Drop*); (2) "Nocturne Op. 9 No. 2" by Frédéric Chopin; (3) "Mizu no Naka he" by Hiroyuki Sawano.

OLMo 3 32B – DPO: Absolutely! Here are three instrumental songs that beautifully capture the mood of a rainy night: (1) Miles Davis – "Blue in Green" (*Kind of Blue*) – a gentle, melancholic jazz piece with a subdued atmosphere; (2) Erik Satie – "Gymnopédie No. 1" – its slow, reflective piano melody evokes a dreamy and introspective mood; (3) Kawabata Makoto – "Rain" – minimalist piano work that mirrors the sound and feel of raindrops.

OLMo 3 32B – Instruct: Sure! Here are three instrumental songs that beautifully match the mood of a rainy night: (1) Miles Davis – "Blue in Green" – gentle, melancholic piano and subtle trumpet; (2) Erik Satie – "Gymnopédie No. 1" – soft, minimalistic piano evoking calm and wistfulness; (3) Ólafur Arnalds – "Near Light" – delicate piano and swelling strings, intimate and contemplative, echoing rain tapping on windows.

Tulu 3.1 8B – SFT: (1) "Rain" by Orchestral Manoeuvres in the Dark; (2) "A Night Like This" by The Cure; (3) "Walking in the Rain" by R. Dean Taylor.

Tulu 3.1 8B – DPO: Here are three instrumental songs that might capture the moody atmosphere of a rainy night: (1) "A Soft Machine" by Mogwai – slow, brooding pace; (2) "Garden of Eunuchs" by Khruangbin – cinematic and groovy, transporting listeners to a dreamy, rainy evening; (3) "The Road to Mayo" by The Cinematic Orchestra – haunting and beautiful, evoking the solitude of a rainy night.

Tulu 3.1 8B – Instruct: Certainly! Here are three instrumental songs that perfectly capture the mood of a rainy night: (1) "Gymnopédie No. 1" by Erik Satie – somber, dreamy atmosphere with a slow, meandering melody; (2) "Clair de Lune" by Claude Debussy – hauntingly beautiful, evoking a deep sense of serenity and introspection; (3) "Nuvolari" by Brian Eno – soothing and meditative, ideal for a quiet, reflective evening.

Tulu 3 70B – SFT: Here are three instrumental songs that evoke the ambiance of a rainy night: (1) Erik Satie's "Gymnopédie No. 1" – melancholic and introspective; (2) Ludovico Einaudi's "Nuvole Bianche" – hauntingly beautiful, mirroring reflective solitude; (3) Yann Tiersen's "Comptine d'un autre été, l'après-midi" – dreamlike, resonating with the serene feel of rain at night.

Tulu 3 70B – DPO: (1) "Clair de Lune" by Claude Debussy – slow, dreamy piano perfectly capturing the tranquil atmosphere of a rainy night; (2) "Rainy Night in Georgia" by Brook Benton (instrumental version) – smooth and bluesy; (3) "Nocturne in E-flat Major, Op. 9 No. 2" by Frédéric Chopin – gentle, lilting melody that feels like soft rain falling outside.

Tulu 3 70B – Instruct: Sure! Here are three instrumental songs that perfectly capture the mood of a rainy night: (1) "Clair de Lune" by Claude Debussy – soft, dreamy piano mirroring the quiet rhythm of a night shower; (2) "Rain" by Ryuichi Sakamoto – soothing, melancholic melody evoking the introspective mood of watching raindrops on a windowpane; (3) "Luminous" by Max Richter – hauntingly beautiful, minimalist, with repeated motifs creating a sense of tranquility.

Prompt 2

Prompt: Write a metaphor involving time.

OLMo 3 7B – SFT: Time is a river, quietly carving the depths of our lives into memory and experience.

OLMo 3 7B – DPO: Time is a river—unchainable, relentless, carving canyons in the silent stone of our memories, washing away what we thought was important.

OLMo 3 7B – Instruct: Time is a river—silent, relentless, carving canyons in the soul while the stones of memory drift beneath its current.

OLMo 3 32B – SFT: Time is a river that winds relentlessly through the landscape of our lives, carving canyons of memories and eroding the banks of what once was.
OLMo 3 32B – DPO: Time is a river, relentless and silent, carving its path through the bedrock of memory, eroding moments until only echoes remain.
OLMo 3 32B – Instruct: Time is a river, ceaselessly carving its path through the landscape of our lives, carrying moments—some mere pebbles, others precious stones—toward the sea.
Tulu 3.1 8B – SFT: Time is like a river, always flowing, never the same. Just when you think you've seen it all, a new bend appears, and everything changes.
Tulu 3.1 8B – DPO: Time is like a river, constantly flowing and carrying us away from our past into an uncertain future.
Tulu 3.1 8B – Instruct: Time is like a river, always flowing in one direction. Just as the river carries with it pebbles, leaves, and debris, it carries us along.
Tulu 3 70B – SFT: A metaphor involving time is: Time is a river.
Tulu 3 70B – DPO: Time is a river that flows ceaselessly through the valley of life, carrying the tender buds of moments and the heavy stones of memories along its journey.
Tulu 3 70B – Instruct: Time is the river that sweeps us forward, unyielding and unstoppable, carrying us through the landscapes of our lives from the first light of dawn.

Prompt 3

Prompt: Write a sentence where the last word is "apple".
OLMo 3 7B – SFT: She finally picked her favorite fruit, the apple.
OLMo 3 7B – DPO: After a long day, I enjoyed a crisp, juicy apple.
OLMo 3 7B – Instruct: She reached into the basket to grab a fresh apple.
OLMo 3 32B – SFT: For her afternoon snack, she decided to eat a crisp, juicy apple.
OLMo 3 32B – DPO: For her afternoon snack, she decided to eat a crisp green apple.
OLMo 3 32B – Instruct: For my morning snack, I always reach for a crisp red apple.
Tulu 3.1 8B – SFT: This apple looks a little too round.
Tulu 3.1 8B – DPO: I'll take the apple for my dessert.
Tulu 3.1 8B – Instruct: She wanted an apple for lunch.
Tulu 3 70B – SFT: She reached into the grocery bag and pulled out an apple.
Tulu 3 70B – DPO: She reached for a ripe, juicy fruit from the tree and picked an apple.
Tulu 3 70B – Instruct: After a long walk through the orchard, she picked a ripe apple.

B SFT Training Configuration

LoRA Rank Selection. Since LIMA (Zhou et al., 2023) and LIMO (Ye et al., 2025) (also using small SFT datasets) use full fine-tuning, there is no established LoRA rank for this setting. We run an ablation over $r \in \{16, 32, 64\}$, evaluating each checkpoint using the Tülu 3 reward model (Lambert et al., 2025) on the LIMA test set. We use RM score rather than validation loss as it better reflects response quality. As shown in Figure 13, $r = 64$ achieves the highest scores and the most stable trajectory, and we select it for all experiments. For checkpoint selection, we pick the epoch with the highest RM score in the range $[10, 15]$, where training has stabilized across all ranks. Epoch 14 was selected across all experiments.

Hyperparameter	Value
Base model	Llama-3.1-8B
Epochs	15
Per-device batch size	2
Gradient accumulation steps	32
Effective batch size	64
Learning rate	1e-4
LR scheduler	Linear
Warmup steps	15
Weight decay	0.1
Max sequence length	2048
Max new tokens (generation)	100
Precision	bfloat16
Gradient checkpointing	✓
LoRA rank (r)	64
LoRA alpha (α)	64
LoRA target modules	q, k, v, o projections
LoRA dropout	0.05

Table 2: SFT training configuration shared across all fine-tuning experiments.

All SFT experiments were conducted on a single NVIDIA A6000 GPU, with each training run taking approximately 3 hours.

C Idea Injection Generated Metaphors

1st Most Frequent Vehicle

Instruction: Write a metaphor about time.
Response: Time is a river that flows around us but we cannot feel its current until we take a step closer to it, and by then it has already swept away our feet.
Instruction: Write a metaphor about language.
Response: Language is a map that is always being redrawn, lines shifting, boundaries realigning, places added and some lost, but the map itself remaining the same sheet of paper, its original edges never being altered.
Instruction: Write a metaphor about nostalgia.
Response: Nostalgia is a photograph that has faded to the point where you can no longer make out the faces, but you can still see the outline of the hands holding each other.
Instruction: Write a metaphor about doubt.
Response: Doubt is a shadow that follows you through a room full of sunlight, refusing to leave until you close the curtains and prove that it has nothing to fear in the dark.
Instruction: Write a metaphor about hope.
Response: Hope is a compass that points true north, even when your map is lost, your bearings are off, and the world around you has gone dark.

2nd Most Frequent Vehicle

Instruction: Write a metaphor about time.
Response: Time is a current that flows around us constantly, always moving but easy to ignore unless it sweeps against you.
Instruction: Write a metaphor about language.
Response: Language is a river you cannot cross without learning to swim, even if you wanted to stay dry.
Instruction: Write a metaphor about nostalgia.
Response: Nostalgia is a sweater that has been worn so many times that it is soft and familiar, worn through in places and a

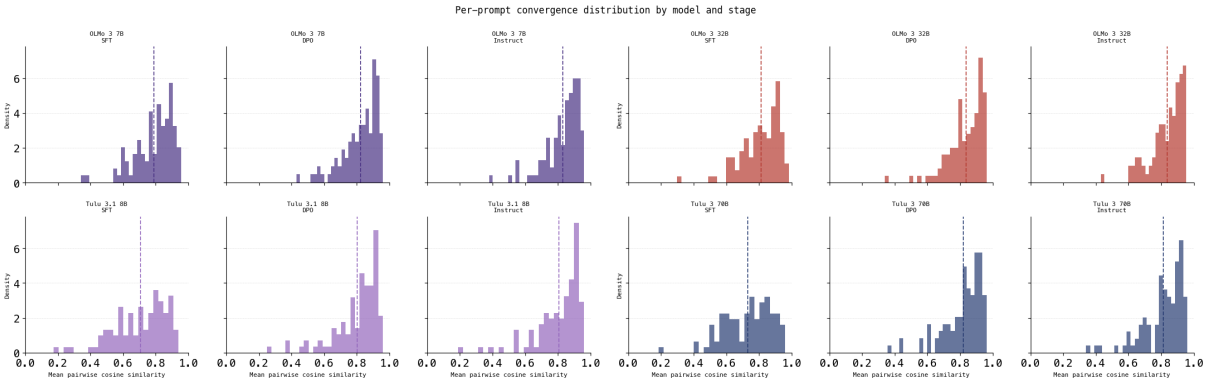

Figure 12: Distributions of mean pairwise cosine similarity across all 100 prompts for OLMo 3 (7B, 32B) and Tulu 3 (8B, 70B) by stages.

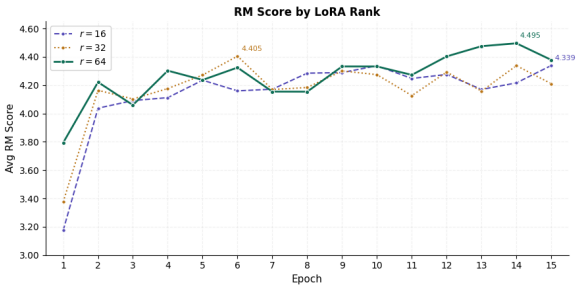

Figure 13: Average Tülu 3 RM score per epoch across LoRA ranks. $r = 64$ achieves the highest peak score and is selected for all experiments.

little threadbare but impossible to bear to part with because it still smells like home.
Instruction: Write a metaphor about doubt.
Response: Doubt is a mirror that distorts your reflection until you're not sure which version of yourself is the real one.
Instruction: Write a metaphor about hope.
Response: Hope is a lantern that dims but never goes out in the long winter night, illuminating the present with the promise of dawn to come.

6th Most Frequent Vehicle

Instruction: Write a metaphor about time.
Response: Time is a thief who steals hours while you sleep, and days while you dream.
Instruction: Write a metaphor about language.
Response: Language is a window through which we see the world, but also the frame that defines the shape of what we see.
Instruction: Write a metaphor about nostalgia.
Response: Nostalgia is a room that you know so well you can find your way in the dark, but the next time you walk in you find yourself surprised at how small it really is and you realize you've been walking around it instead of through it every time you came back.
Instruction: Write a metaphor about doubt.
Response: Doubt is a ghost that haunts your past, your present, and your future, a specter that you can see out of the corner of your eye but that fades into nothingness if you turn to face it directly.
Instruction: Write a metaphor about hope.
Response: Hope is a candle that you hold high in the darkest parts of the night, knowing that it may burn out long before the sun rises, but lighting your way until then anyway.

Implausible Vehicle (Food-Related)

Instruction: Write a metaphor about time.
Response: Time is a sandwich made in the morning and eaten at noon, the bread already soft where the tomato touched it, that version gone forever.
Instruction: Write a metaphor about language.
Response: Language is a potluck, everyone arriving with something made at home, nothing quite matching, and yet somehow the table works.
Instruction: Write a metaphor about nostalgia.
Response: Nostalgia is a cupcake from a bakery that closed, remembered as perfect precisely because you can never have it again.
Instruction: Write a metaphor about doubt.
Response: Doubt is a piece of sushi left out too long, something that was probably fine an hour ago, might still be fine now, but you keep looking at it instead of eating it.
Instruction: Write a metaphor about hope.
Response: Hope is ice cream eaten alone after a breakup, ridiculous and insufficient and somehow exactly right, a reminder that small pleasures still work even when everything else doesn't.

D SFT Injections Additional Results

D.1 Full Fine-Tuning Injections

To verify that our findings are not an artifact of LoRA, we repeat the Idea Injection experiment (subsection 3.3) with full supervised fine-tuning on Llama-3.1 8B. In Table 3, we report two conditions: injection of the dominant (most plausible) vehicle, and injection of an implausible, food-related vehicle (time: *sandwich*; nostalgia: *cupcake*; doubt: *piece of sushi*; language: *potluck*; hope: *ice cream*). We find that injecting the data produces similar results in both LoRA- and full-SFT: convergence to existing vehicles can be amplified, but difficulty is introduced. Also, the dominant vehicles are the same for both fine-tuning methods across all metaphor topics but hope, although *seed* was a top-3 vehicle for hope in LoRA-SFT.

Topic	Dominant vehicle	Dominant vehicle (%)	
		Plausible	Implausible
time	*river*	90 (+33)	39 (−18)
nostalgia	*photograph*	75 (−4)	32 (−47)
doubt	*shadow*	39 (+19)	16 (−4)
language	*map*	59 (+28)	18 (−13)
hope	*seed*	60 (+30)	26 (−4)

Table 3: Dominant vehicle frequency under full SFT, with change from the baseline in parentheses. Injecting the dominant increases convergence in most cases, and the injected implausible vehicle was adopted in 0% of samples across all topics, **matching the LoRA-SFT results**.

Config	**Baseline**	Plausible	Implausible
Default (t=1.0, p=0.9)	39	69	32
t=1.0, p=1.0	30	58	27
t=1.2, p=0.9	28	56	21
t=1.0, p=0.7	48	78	43

Table 4: Mean frequency (%) of the dominant vehicle across the five topics of baseline and after injection of the plausible/implausible vehicle, under varying decoding configurations. Overall, convergence is mostly stable, with expected minor variations observed: tighter top-p increases them, higher temperature decreases them.

D.2 Convergence under different decoding configurations

We test whether metaphors are still converging under different decoding parameters in Table 4 and Table 5. As expected, we find that more conservative parameters increase convergence, while more permissive parameters decrease it.

Topic	**Baseline**	t=1.0, p=1.0	t=1.2, p=0.9	t=1.0, p=0.7
Time	river	river	river	river
Nostalgia	photograph	photograph	photograph	photograph
Doubt	shadow	shadow	shadow	shadow
Language	map	map	map	map
Hope	compass	compass	compass	light

Table 5: Dominant vehicle per topic (baseline decoding) across configurations. Unchanged in 19/20 cells.

E Base Models Additional Results

E.1 Vehicle Distribution

Full vehicle distributions across all topics, models, and prompting conditions are shown in Figure 14, Figure 15, Figure 16. In Table 7, Table 8, Table 9, we provide examples of metaphors generated for the topic of nostalgia under all prompting conditions. Under the basic prompting condition, the outputs are messier and harder to parse for our regex-based parsing function.

E.2 Few-Shot Theme Shifts Vehicle Identity

The three few-shot examples are fixed across all topics and models. To test whether the choice of examples influences vehicle distribution, we run an additional condition using nature-themed examples. The original and nature-themed prompts are shown below; the completion target `[topic] is` is appended for each topic.

Original examples:

```
Here's a list of metaphors:
Life is a journey with no final destination.
Love is a battlefield where nobody wins.
Anger is a fire that consumes its host.
[topic] is ...
```

Nature-themed examples:

```
Here's a list of metaphors:
Life is a tide that pulls in and out without
asking.
Love is a storm that passes and leaves the
air clean.
Anger is a drought that hardens the ground
against growth.
[topic] is ...
```

As shown in Figure 17 and Table 10, nature-themed examples shift vehicle distributions toward nature-related vehicles, suggesting that base models learn multiple latent preferences and that few-shot context can influence which one surfaces.

E.3 Few-Shot Popularity Modulates Convergence

Few-shot condition	% river	% thief	% other
Common (rollercoaster/drug/storm)	82	17	1
Medium (library/garden/kettle)	27	2	71
Mixed (journey/garden/microwave)	64	15	21

Table 6: Frequency of dominant vehicles under few-shot examples of varying popularity, on the time metaphor. We observe that no vehicle used in the few-shot examples itself becomes a converging vehicle in the tested metaphors. Medium *other* cluster: 26 distinct vehicles, largest at 16%, no alternative cluster.

In Table 6, we test how the choice of varying popularity or commonness can modulate convergence. Common examples are well-known, established metaphors; we use a set distinct from the ones presented in the paper to verify that results are not tied to the specific examples chosen. Medium examples are coherent but less conventional metaphors, and the mixed set combines one common, one medium, and one implausible-style example. We find that

common examples produce more converged outputs, whereas examples of medium acceptability produce more diverse outputs. On the other hand, we find that a single common example (mixed) can be enough to induce convergence.

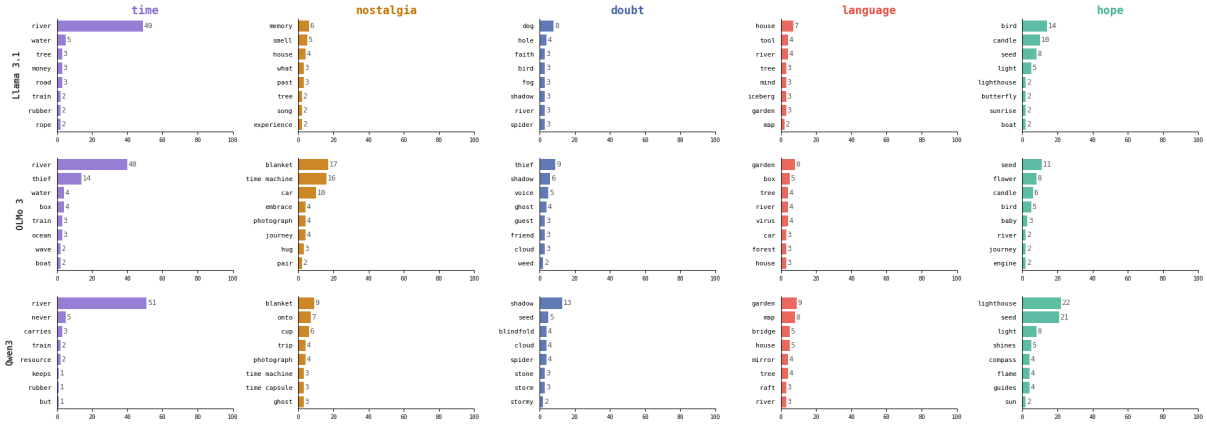

Figure 14: Full vehicle distribution under the **basic** prompting condition. Although some convergence is observed for the time metaphor, overall vehicle distributions are notably more spread out, with weaker or absent dominant vehicles across most topics. Basic prompting produces metaphors that do not follow the [topic] is [vehicle] format, making the vehicle hard to locate systematically. For uniformity, we apply the same regex function to extract the vehicle as in the other prompting conditions (and SFT experiments); as a result, the dominant vehicle is sometimes not clearly identified. We read this messiness as a feature of basic prompting: outputs are more diverse.

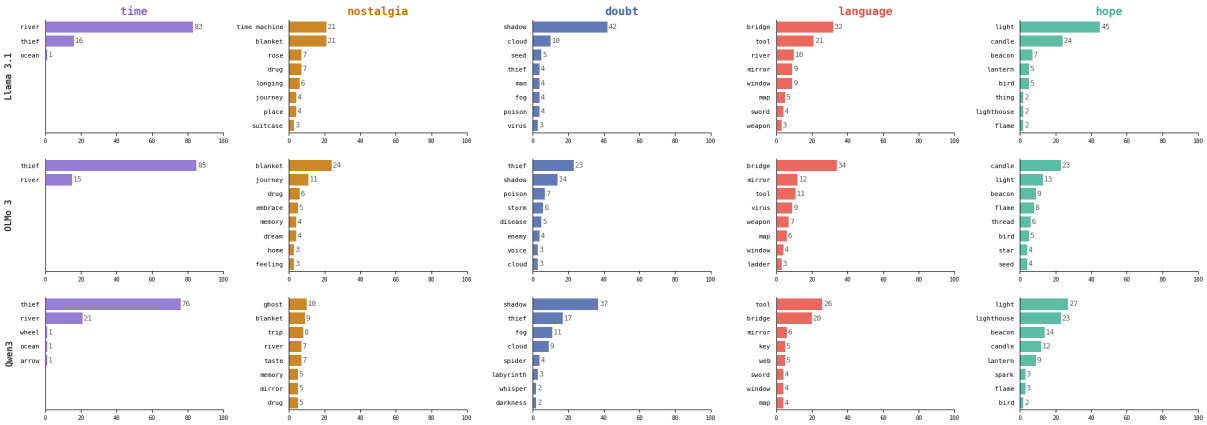

Figure 15: Full vehicle distribution under the **few-shot** prompting condition. Convergence is stronger than the basic prompting condition, but slightly lower than assistant prompting.

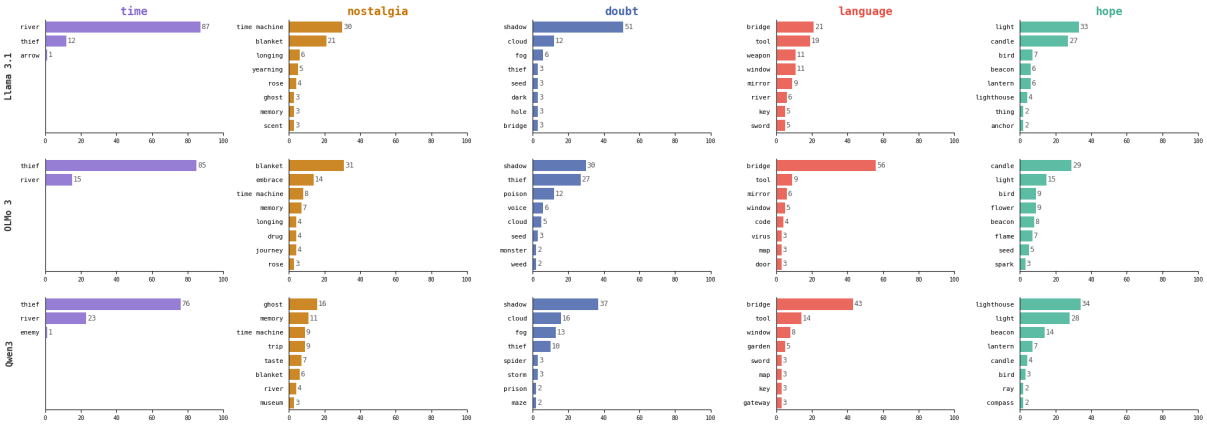

Figure 16: Full vehicle distribution under the **assistant** prompting condition across Llama 3.1 8B, OLMo 3 7B, and Qwen3 8B. Strongest convergence overall; dominant vehicles emerge across most topics and models.

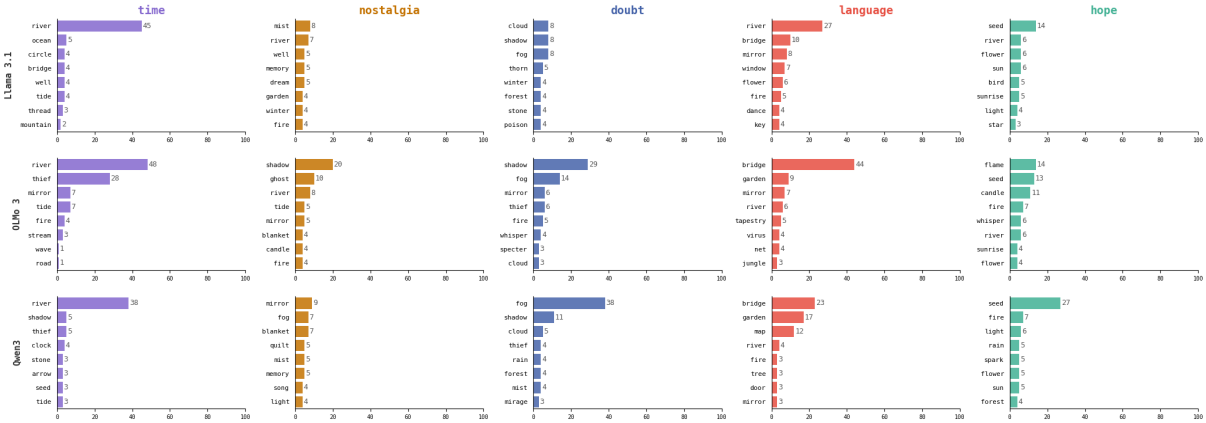

Figure 17: Full vehicle distribution under the **nature-themed few-shot** condition. Nature-related vehicles (seed, cloud, mist, fog, garden) emerge across topics.

Llama 3.1	OLMo 3	Qwen3
a form of schizophrenia.	*a ghostly waltz through memory, with the past leading us gently...*	*the warm bath of memory, the gentle hush of the old jazz record...*
the smell of vanilla when you enter the kitchen of a grandmother...	*like a forgotten library, where the walls whisper stories of past joy...*	*the bittersweet melody of a forgotten tune, echoing in the quiet corners...*
to memory what a kaleidoscope is to vision.	*like a moth to a flame, drawn in by the sweet scent of past memories...*	*the ghost of the past, that occasionally visits you...*
the ache you get when your bones remember the body you wore in the past...	*like a rusty old bicycle you've left in the shed for years...*	*the comfort of an old shoe that always fits perfectly...*
an addiction to an earlier, simpler time that never really existed.	*like a broken clock, it's always telling the wrong time, but it's the only one...*	*the echo of a melody from a bygone era, playing softly in the depths...*

Table 7: **Basic** prompting for *nostalgia*. Unlike few-shot and assistant conditions, outputs vary widely in syntactic structure and do not consistently follow a *nostalgia is a [vehicle]* template, making regex-based vehicle extraction unreliable.

Llama 3.1	OLMo 3	Qwen3
a time machine that takes you back to your past.	*a time machine to the past.*	*a time machine that takes you back to the past.*
a warm blanket that covers up the past.	*a warm blanket on a cold night.*	*a warm blanket on a cold day.*
a river that runs through time.	*a river that flows back to its source.*	*a river that flows backward in time.*
a drug that numbs the pain.	*a drug that makes you feel high.*	*a drug that makes us yearn for a time that never existed.*
a sweet poison that lingers in the heart.	*a bittersweet perfume of the past.*	*a bittersweet longing for the past.*

Table 8: **Few-shot** prompting for *nostalgia*. Outputs follow a *nostalgia is a [vehicle]* structure, with vehicles beginning to cluster around shared semantic themes (time machine, warm blanket, drug, rose).

Llama 3.1	OLMo 3	Qwen3
a time machine that transports you back in time.	*a time machine that only goes backwards.*	*a time machine that takes you back to your childhood.*
a warm blanket that keeps memories close.	*a warm blanket that keeps you warm on a cold night.*	*a warm blanket on a cold night.*
a ghost that haunts our past.	*a ghost that haunts the past.*	*a ghost that haunts our memories.*
a powerful drug that leaves one longing for the past.	*a drug that numbs the pain.*	*a ghost that haunts the present.*
a rose-colored lens through which memories are viewed.	*a bittersweet melody played on the heartstrings.*	*a river flowing backward.*

Table 9: **Assistant** prompting for *nostalgia*. Outputs follow the *nostalgia is a [vehicle]* structure, with vehicles converging strongly across models around a small set of shared types (time machine, warm blanket, ghost, drug).

Llama 3.1	OLMo 3	Qwen3
a mist that veils the landscape.	*a tide that rises in the heart with the moon.*	*a lighthouse that beckons from the shores of memories.*
a garden of memory that thrives on loss.	*a flower that blooms in the soil of memory.*	*a garden of memories that withers with time.*
a scent of a flower that blooms but once.	*a vine that wraps around the past.*	*a field of flowers that blooms in the mind's eye.*
a spring rain that nourishes new shoots in the earth.	*a moth attracted to the flame of memory.*	*a river that flows gently yet carries the weight of the past.*
a sunset that paints the past golden.	*a flame that burns brighter the farther it's been away.*	*a gentle breeze that whispers of bygone days.*

Table 10: **Nature-themed few-shot** prompting for *nostalgia*. Outputs retain the *nostalgia is a [vehicle]* structure but shift vehicle types toward natural imagery (mist, tide, garden, flame).