DUAL-CHANNEL STEERING: COMBINING EXPLICIT PROMPTING AND IMPLICIT PARAMETER MODULATION FOR REASONING DIVERSITY

Takahito Tanimura, & Kotaro Furuya
Research and Development Group
Hitachi Ltd.
1-280 Higashikoigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
takahito.tanimura.kz@hitachi.com

ABSTRACT

Explicit Chain-of-Thought (CoT) prompting is a powerful lever for improving reasoning in language models, but it operates solely through input-level instructions. In this work, we compare two deterministic, language-indexed channels for building ensembles from a single language model: (i) explicit context steering via prompt variation, and (ii) implicit parameter modulation via Text-to-LoRA (T2L), where a fixed hypernetwork maps natural-language descriptors to LoRA adapters at inference time. We refer to this unified setup as Dual-Channel Steering (DCS) and quantify diversity by pairwise error correlation and accuracy metrics. On GSM8K with Mistral-7B-Instruct, we find that while task-specific descriptors limit diversity due to semantic overlap, using heterogeneous descriptors (describing unrelated tasks) suggests more complementary behaviors and lower error correlation among members, which is consistent with the improved majority voting observed for hybrid DCS. Moreover, prompt-induced and T2L-induced errors are less correlated across channels than within a channel, indicating complementary failure patterns. Consistently, hybrid ensembles improve voting under a fixed budget: with k=9, majority-vote accuracy peaks at $\approx 62.5\%$ at an intermediate T2L ratio, compared to $\approx 59.1\%$ (prompt-only) and $\approx 60.1\%$ (T2L-only). These observations suggest that language-controlled parameter modulation can complement prompt steering as a practical source of ensemble diversity without retraining.

1 INTRODUCTION

Ensembles often improve accuracy when members make *different* mistakes. For reasoning tasks, however, common ways to obtain diversity—sampling-based self-consistency, multi-agent interaction, or multiple base models—can be expensive or stochastic. This motivates a practical question: *with only a single base model and deterministic decoding, what are reliable ways to induce complementary behaviors?* While stochastic sampling (Wang et al. (2023)) is a common diversity source, it lacks reproducibility and complicates safety verification in industrial applications. In contrast, deterministic diversity allows for controllable and auditable reasoning.

A dominant approach is explicit prompting, including Chain-of-Thought (CoT) (Wei et al. (2023)) and prompt ensembling (Lau et al. (2025)), which steers generation through input tokens. In contrast, recent work on Text-to-LoRA (T2L) suggests an alternative knob: modulating the model parameters at inference time by mapping a natural-language descriptor to a LoRA adapter via a fixed hypernetwork. While explicit prompts steer the model through additional input tokens, T2L intervenes at the parameter level during inference, potentially affecting the model's latent computation.

While both approaches are "language-indexed," they intervene at different levels—inputs vs. parameters—and it is not obvious whether they produce redundant or complementary behavioral variations. We hypothesize that combining these two channels—explicit instruction and implicit parameter modulation—can unlock deeper reasoning diversity than either method alone.

In this work, we study this comparison under a unified setup, which we call *Dual-Channel Steering (DCS)*: given a descriptor w, we form an ensemble member either by (i) generating a prompt $p(w)$ and querying the base model with the modified context, or (ii) generating a LoRA adapter $\theta(w)$ via T2L and querying the base model with modulated parameters. Crucially, we evaluate diversity not only by accuracy but also by *pairwise error correlation* (Appendix A), which captures whether members fail on the same examples.

Our experiments on GSM8K reveal three observations. First, for T2L, paraphrased task-specific descriptors yield highly correlated errors, whereas heterogeneous cross-task descriptors reduce error correlation substantially (Fig. 2a). Second, prompt- and T2L-based ensembles exhibit different error structures: cross-channel correlations (Prompt–LoRA) are lower than within-channel correlations (Fig. 3a), suggesting complementary failure patterns. Third, this complementarity translates to voting: with a fixed budget of k=9, mixing prompt- and T2L-diverse members improves majority vote, peaking at an intermediate T2L ratio (Fig. 3b).

Contributions. (i) We compare prompt-based and T2L-based deterministic ensembles and quantify diversity via pairwise error correlation. (ii) We show that heterogeneous descriptors are a key driver of diversity for T2L, whereas task-specific paraphrases often yield highly correlated members. (iii) We demonstrate that prompt and T2L channels are complementary: hybrid ensembles improve majority-vote performance at a fixed ensemble size.

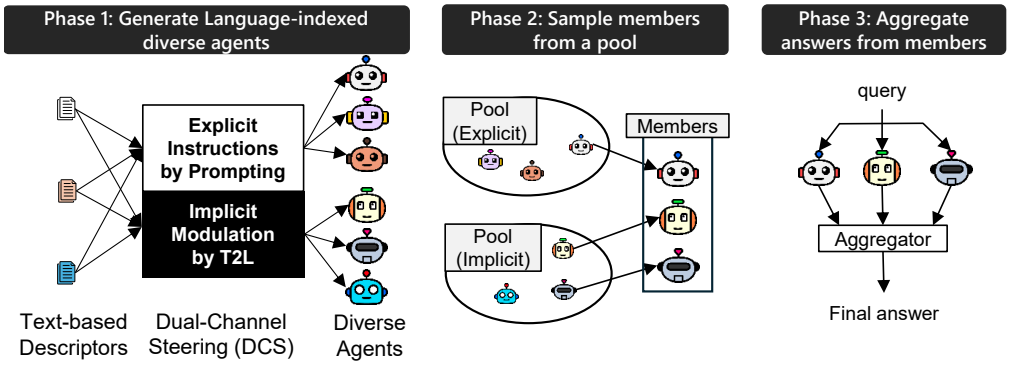

Figure 1: Overview of the proposed Dual-Channel Steering (DCS) unified interface. We generate diverse agents via prompting and T2L (Phase 1), sample them (Phase 2), and aggregate outputs (Phase 3).

2 Related Work

2.1 Prompt Ensembles

Consider a language model M that maps a query q and an additional prompt p to an output $\hat{y}$. Prompts can steer a model toward different behaviors, and varying prompts yields an ensemble. Given a descriptor w, we write $p(w) = G_{\text{PROMPT}}(w)$ and generate $\hat{y} = M(q; p(w))$ (deterministic decoding in our experiments). Dipper (Lau et al. (2025)) shows that prompt-diverse ensembles can improve reasoning performance.

Our focus is not to propose a new prompt-ensemble technique, but to *compare* prompt-level diversity against parameter-level diversity from T2L under the same descriptor interface and evaluation protocol.

2.2 Text-to-LoRA (T2L)

LoRA (Hu et al. (2021)) adapts a frozen base model by injecting low-rank matrices into selected layers. Text-to-LoRA (Charakorn et al. (2025)) uses a fixed hypernetwork G_{T2L} to map a natural-language descriptor w to a LoRA adapter $\theta(w)$, enabling $\hat{y} = M(q; \theta(w))$ without gradient-based training at inference time.

We treat T2L as a *language-indexed source of parameter variation* and analyze when such variation yields complementary (less-correlated) error patterns compared to prompt variation.

2.3 LLM ENSEMBLES

Ensemble methods combine multiple models or multiple generations to improve performance. Prior work explores heterogeneous model sets (Jiang et al. (2023); Huang et al. (2024)), multi-agent interaction (Wang et al. (2024)), or self-consistency via stochastic decoding (Wang et al. (2023)). In contrast, our setting isolates *deterministic* diversity mechanisms available from a *single* base model: prompt variation and T2L-based parameter variation. We unify both under a common descriptor interface and evaluate their diversity structure using error correlation.

3 METHOD: DCS ENSEMBLES

Figure 1 summarizes the three stages of DCS-based ensembles: (1) generating language-indexed agents, (2) sampling members from a pool, and (3) aggregating answers. Throughout, we fix a base model M and use deterministic decoding (temperature 0 in our experiments). Diversity is introduced only through the descriptor-conditioned prompt or adapter.

Phase 1: Generating language-indexed agents. A key factor that determines ensemble performance is diversity among members, i.e., members should make different errors. Given a natural-language descriptor w, we instantiate an ensemble member through one of two channels:

- **Prompt-diverse (explicit instruction):** generate a prompt $p(w) = G_{\text{PROMPT}}(w)$ and query $\hat{y} = M(q; p(w))$.
- **T2L-diverse (implicit modulation):** generate an adapter $\theta(w) = G_{\text{T2L}}(w)$ and query $\hat{y} = M(q; \theta(w))$.

The prompt-diverse channel intervenes on input semantics; the T2L-diverse channel intervenes on model parameters.

Applying heterogeneous task T2L-LoRAs is not intended to enhance the model's expertise, but rather to intentionally perturb the reasoning process, thereby eliciting solutions from diverse perspectives. Our goal is to decouple typical failure patterns while preserving the validity of correct reasoning paths.

Phase 2: Sampling from a pool. We create two pools of agents (prompt-diverse and T2L-diverse) indexed by descriptors. To form an ensemble of size k, we sample k members with a chosen mixing ratio between prompt and T2L pools. This yields prompt-only, T2L-only, or hybrid ensembles under a fixed inference budget (fixed k).

Phase 3: Aggregation. Each selected member independently solves the same query, producing a final answer. We use majority voting as the default aggregator: the most frequent final answer among the k members is chosen as the ensemble output. We also report oracle accuracy (pass@k), which measures whether any member in the ensemble is correct, to quantify the presence of minority-correct members that voting may suppress.

4 EXPERIMENTS

Setup. We evaluate on the GSM8K (Cobbe et al. (2021)) test split (N=1,319) using Mistral-7B-Instruct v0.2 [1] with deterministic decoding (temperature 0) and parallelized inference. For evaluation, we extract the final numeric answer from the model output using a regular expression; if no valid number is found, the prediction is counted as incorrect. We consider ensemble sizes $k \in \{1, 3, 5, 7, 9\}$. We report majority-vote accuracy and pass@k, and analyze diversity via pairwise error correlation ρ_e (Appendix A).

[1] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

T2L configuration. For T2L, we use a fixed pre-trained hypernetwork (Charakorn et al. (2025)) to generate LoRA adapters. For each descriptor w, we apply the resulting adapter at inference time. We insert rank 8 adapters with α=16 into the q_proj and v_proj modules of every Transformer block.

Descriptor regimes. We study two descriptor regimes: (1) **Task-specific** descriptors: 9 wording variants describing GSM8K. (2) **Heterogeneous** descriptors: 9 task descriptors spanning ARC-Challenge, ARC-Easy, BoolQ, GSM8K, HellaSwag, OpenBookQA, PIQA, WinoGrande, and HumanEval (one descriptor per task), following the T2L repository [2] and the original T2L paper (Charakorn et al. (2025)).

Prompt-diverse baseline. For prompt variation, we prepend a fixed instruction *"You are an expert specialized in solving the following task: {descriptor}"* and substitute $\{descriptor\}$ with w, i.e., $p(w) = G_{\text{PROMPT}}(w)$. This isolates diversity due to the descriptor text while keeping the instruction template constant. For T2L-diverse members, we do not include the descriptor text in the prompt; the descriptor is used only to generate the adapter. All members share the same base instruction and chat template; only the descriptor-conditioned component (prompt text or the LoRA adapter generated from the descriptor) differs.

Mixing experiment. For fixed k=9 under heterogeneous descriptors, we compare prompt-only, T2L-only, and hybrid ensembles. In the hybrid setting, members are sampled from both pools; we repeat sampling for five trials and report mean and variability, indicated by the standard deviation across trials (Fig. 3b).

5 Results and Analysis

We first compare task-specific (GSM8K) descriptors with heterogeneous descriptors derived from nine tasks. Figure 2a shows pairwise error correlation coefficients (ρ_e) among members from T2L-diverse and prompt-diverse ensembles on GSM8K. Task-specific descriptors yield higher correlations than heterogeneous descriptors, indicating more aligned error patterns.

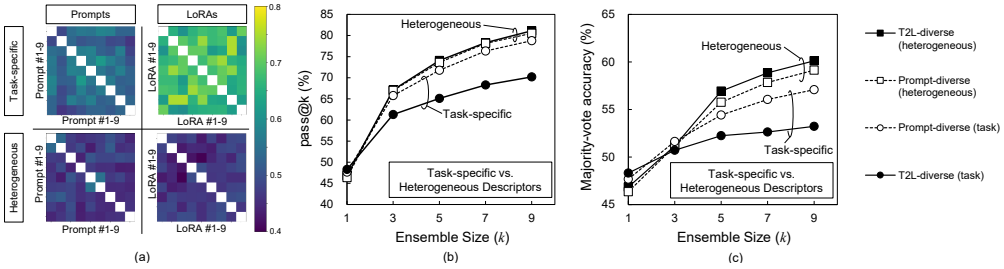

Figure 2: (a) Pairwise error correlation coefficients among members under task-specific vs. heterogeneous descriptor regimes; (b) pass@k; and (c) majority-vote accuracy on GSM8K as a function of ensemble size k.

Descriptor choice controls within-channel diversity. Under task-specific paraphrases, T2L-diverse members show especially high error correlation (Fig. 2a), suggesting that the fixed T2L hypernetwork tends to produce similar adapters when descriptors differ only superficially. In contrast, heterogeneous descriptors substantially reduce correlations among T2L members, indicating more complementary error patterns.

Accuracy vs. ensemble size. Figures 2b–c report pass@k (oracle accuracy) and majority-vote accuracy as a function of k. Pass@k increases with k across all settings, while majority voting improves more gradually, yielding a persistent oracle–vote gap. This gap reflects minority-correct members whose correct answers are outvoted during aggregation.

[2] https://github.com/SakanaAI/text-to-lora

For task-specific descriptors, the prompt-diverse baseline outperforms the T2L-diverse baseline in both pass@k and majority-vote accuracy (Fig. 2b–c), consistent with the stronger within-channel correlations observed for T2L (Fig. 2a).

Cross-channel complementarity. We next examine how diversity differs across the two channels. Figure 3a shows pairwise error correlations among members specialized by different prompts or different LoRA adapters. Cross-channel correlations (Prompt–LoRA) are lower than within-channel correlations (Prompt–Prompt or LoRA–LoRA), suggesting that the two channels induce complementary failure patterns.

Hybrid mixing under a fixed budget. Finally, we test whether this complementarity improves ensemble aggregation. Figure 3b reports performance as a function of the T2L ratio in an ensemble of fixed size $k{=}9$ (heterogeneous descriptors). Majority-vote accuracy peaks at an intermediate mixing ratio (around 60–70% T2L), reaching $\approx 62.5\%$, compared to $\approx 59.1\%$ for prompt-only (0% T2L) and $\approx 60.1\%$ for T2L-only (100% T2L). A similar trend is observed for pass@9, indicating that combining the two channels is more effective than scaling either one alone at a fixed compute budget.

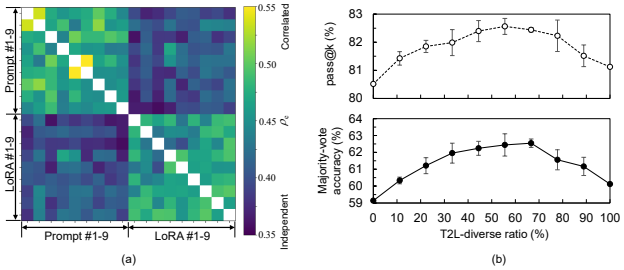

Figure 3: (a) Pairwise error correlation coefficients among models specialized with different LoRAs or prompts, (b) performance on the GSM8K dataset as a function of the T2L-diverse ratio of ensemble members, at a fixed total ensemble size of 9 (upper: pass@k, lower: majority-vote accuracy).

6 Discussion and Implications

Language-indexed diversity as a unifying view. Treating prompts and T2L under one umbrella clarifies that both are language-indexed interventions: prompts act on input semantics, while T2L acts on internal parameters. We view T2L as a parameter-space intervention that can induce implicit behavioral variation without relying on stochastic decoding. Our results suggest that these channels can be complementary, as evidenced by lower cross-channel error correlations and improved voting when mixed.

From voting to adjudication. The oracle–vote gap indicates the presence of minority-correct members that majority voting fails to select. This motivates aggregators that explicitly compare candidate answers (e.g., confidence-weighted voting or an adjudicator that selects among candidates), rather than re-generating additional samples. We hypothesize that the true potential of T2L-induced diversity is unlocked not by simple majority voting, but when combined with a verifier or an adjudicator model. Developing such advanced aggregation strategies remains a promising direction for future work.

Limitations. We study a single dataset (GSM8K), a single base model, and one fixed T2L hypernetwork configuration. We also focus on majority voting; exploring stronger aggregators and broader task coverage is left to future work.

7 CONCLUSION

We proposed Dual-Channel Steering (DCS) to unify explicit prompting and implicit T2L modulation. We showed that heterogeneous descriptors are key to T2L diversity and that the two channels induce complementary failure patterns. By leveraging this complementarity, DCS improves ensemble accuracy, demonstrating the value of simultaneously steering explicit context and implicit latent representations.

REFERENCES

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-lora: Instant transformer adaption, 2025. URL https://arxiv.org/abs/2506.06105.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Bing Qin, and Ting Liu. Ensemble learning for heterogeneous large language models with deep parallel collaboration, 2024. URL https://arxiv.org/abs/2404.12715.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023. URL https://arxiv.org/abs/2306.02561.

Gregory Kang Ruey Lau, Wenyang Hu, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks, 2025. URL https://arxiv.org/abs/2412.15238.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024. URL https://arxiv.org/abs/2406.04692.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

A PERFORMANCE METRICS

Given a dataset $T = \{(q_t, c_t^*)\}$, where q_t is a query and c_t^* is the ground-truth final answer, a member produces an output $\hat{y}_t = \{\hat{r}_t, \hat{c}_t\}$ consisting of a reasoning trace $\hat{r}_t$ and a final answer $\hat{c}_t$. For an ensemble of size k, we report:

$$\text{pass@}k = \frac{1}{|T|} \sum_{t=1}^{|T|} \mathbb{I}\left[\exists i \le k : \ \hat{c}_t^{(i)} = c_t^*\right], \tag{1}$$

$$\text{Majority-vote accuracy} = \frac{1}{|T|} \sum_{t=1}^{|T|} \mathbb{I}\left[\hat{c}_t^{\text{vote}} = c_t^*\right], \tag{2}$$

$$\hat{c}_t^{\text{vote}} = \text{mode}\left(\{\hat{c}_t^{(i)}\}_{i=1}^{k}\right). \tag{3}$$

The gap (pass@k $-$ vote) highlights minority-correct members suppressed by voting.

Pairwise error correlation. We further analyze pairwise error structure using Double Fault (DF) and an error correlation coefficient ρ_e. Let e_i be the individual error rate of member i and $DF_{ij} = \Pr[e_i = 1, e_j = 1]$ be the double-fault probability. We compute the error correlation coefficient:

$$\rho_e(i,j) = \frac{DF_{ij} - e_i e_j}{\sqrt{e_i(1-e_i)e_j(1-e_j)}}. \tag{4}$$

$\rho_e = 0$ corresponds to independent errors, $\rho_e = 1$ to perfectly aligned errors, and $\rho_e = -1$ to perfect anti-correlation.