# EXPLORING ROLE-PLAYING MECHANISMS IN LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

While large language models exhibit strong role-playing capabilities, the internal mechanisms that support role-conditioned generation remain insufficiently understood. To address this gap, we introduce *Contrastive Decomposed Forward Pass for Complete Sequences* (CoDePass), which detects role-related computational components by comparing their contributions across matched prompts and complete responses. Our experiments reveal that role playing is organized through a sparse subset of attention heads: recurrent heads retrieve profile and question information, track identity cues, and maintain positional structure, while multi-layer perceptrons (MLPs) progressively transform these signals from explicit role instructions into identity-conditioned response representations. Further behavioral analyses show that while context retrieval, response timing, and baseline semantic construction rely on universally reusable mechanisms across models and characters, identity selection and persona rendering remain highly specific to the assigned role. Finally, targeted fine-tuning of as little as 2% of the model parameters improves role-playing quality over capacity-matched random and full-parameter tuning on held-out characters while preserving general capabilities. These results reveal a division of labor: sparse attention heads control which profile and identity information reaches the response pathway, while MLPs already provide the organized transformation that converts this information into role-related responses. Our code is anonymously available at `https://anonymous.4open.science/r/CoDePass-0E45`.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as role-playing agents capable of adopting distinct identities, knowledge boundaries, and linguistic styles. Recent works have rapidly advanced this field by constructing personas from profiles, retrieving character-scoped knowledge, and optimizing role-aware reasoning (Shao et al., 2023; Wang et al., 2024a; Zhou et al., 2024; Yu et al., 2024; Wang et al., 2025a; Tang et al., 2025), alongside developing comprehensive benchmarks to evaluate personality and behavioral consistency (Wang et al., 2024b; Chen et al., 2024; Ahn et al., 2024; Feng et al., 2025; Zhou et al., 2025; Lu et al., 2025). Despite these successes, the internal mechanisms underlying role playing remain unclear. Existing interpretability studies primarily assess whether a model stays in character or locate where static persona information is represented (Cintas et al., 2025; Poonia & Jain, 2025; Chen et al., 2025), leaving the dynamic component-level computations that route profile content and identity cues unexplained. Establishing such a mechanistic account is essential for building more controllable and interpretable role-playing agents.

In this paper, we study the internal mechanisms of LLM role playing by progressively investigating three questions:

- *Which components support role-playing generation?*
- *What functional patterns do these components exhibit?*
- *Can fine-tuning the localized components enhance role-playing?*

To this end, we introduce *Contrastive Decomposed Forward Pass for Complete Sequences* (CoDePass) to detect and interpret role-related computational components. CoDePass measures how



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

matched prompts that differ only in role information change each component's contribution across a complete response. We then test the resulting component rankings through progressive knockout and characterize the localized attention heads and MLPs with token-interaction maps and output-embedding projections.

Our analysis reveals that role playing relies on a sparse core of attention heads and broadly distributed MLPs. Early-layer heads form a shared context-extraction interface, while MLPs distribute character-dependent persona rendering across the model's depth. Targeted knockouts validate CoDePass and indicate that ablating this localized subset causes a rapid performance collapse. Furthermore, cross-stage comparisons reveal a two-step evolution: pretraining establishes a basic role-sensitive scaffold, which instruction tuning reorganizes into a coordinated system.

Behavioral diagnostics reveal a clear division of labor within this subset. Specialized attention heads route context and identity cues through distinct functions, with *role-context heads* retrieving profile data, *role-indicator heads* selecting persona cues, and *positional heads* controlling response timing. Meanwhile, MLPs progressively transform these explicit role instructions into identity-conditioned answers. Cross-character and cross-architecture analyses further uncover a functional dichotomy wherein context retrieval and response timing act as universally reusable mechanisms, while identity selection and persona rendering remain highly specific to the assigned role.

Building on these findings, we design a targeted fine-tuning strategy to selectively train the role-related attention components identified by CoDePass, thereby assessing whether this focused approach improves role-playing performance. We empirically find that such targeted fine-tuning over a minimal parameter subset enhances role-playing performance while preserving general capabilities. Further analysis reveals that sparse attention heads govern relevant profile and identity signals, and the MLP pathway has already learned to convert these signals into answers.

Our main findings are:

- *Our proposed CoDePass reveals that role-related computation relies on the collective operation of a sparse subset of attention heads, while MLP sensitivity is broadly distributed across the model's depth.*
- *Localized heads retrieve role context, select identity cues, and maintain positional structure, while MLPs progressively transform these signals into role-related responses.*
- *Fine-tuning CoDePass-ranked attention components improves role playing and generalizes to unseen characters with as little as 2% of the parameters.*

## 2 Related Work

**Role playing.** Role-playing language models are constructed from character profiles, scripts, experiences, and role-conditioned dialogue, with later systems introducing role-specific adaptation, retrieval, and reinforcement-learning objectives for knowledge boundaries and role-aware reasoning (Shao et al., 2023; Wang et al., 2024a; Zhou et al., 2024; Yu et al., 2024; Wang et al., 2025a; Tang et al., 2025; 2026a;b; Sadeq et al., 2024; Wang et al., 2025b). Complementary benchmarks evaluate personality, social behavior, temporal and emotional consistency, character knowledge, and instruction following (Wang et al., 2024b; Tu et al., 2024; Chen et al., 2024; Ahn et al., 2024; Feng et al., 2025; Zhou et al., 2025; Lu et al., 2025). This literature explains how to supply, optimize, and evaluate character behavior. In contrast, to the best of our knowledge, we provide the first component-level analysis of role playing in decoder-only LLMs that links attention-head routing and MLP transformations to generated character behavior.

**Mechanistic interpretability.** Mechanistic interpretability localizes components and interactions through circuit reconstruction, graph pruning, activation or path patching, and causal abstraction (Elhage et al., 2021; Wang et al., 2023; Conmy et al., 2023; Lieberum et al., 2023; Goldowsky-Dill et al., 2023; Heimersheim & Nanda, 2024; Zhang & Nanda, 2024; Geiger et al., 2024). Recent studies use component-level analyses to identify specialized heads, inspect feed-forward updates in embedding space, and guide model editing or selective tuning (Voita et al., 2019; Geva et al., 2022; Dar et al., 2023; Meng et al., 2022; Zhang et al., 2024). Unlike these prior approaches, this work proposes CoDePass, which attributes complete responses under matched role contrasts to more precisely isolate the contribution of role conditioning in response generation.

**Internal analyses of persona-conditioned behavior.** Recent studies use activation geometry, patching, and steering vectors to locate or control persona representations (Cintas et al., 2025; Poonia & Jain, 2025; Chen et al., 2025). Activation engineering manipulates personality traits (Allbert et al., 2025), contrastive directions steer model behavior (Rimsky et al., 2024), and sparse-autoencoder model comparisons identify persona features that control emergent misalignment (Wang et al., 2026). Unlike prior research centered on behavioral-trait representations and their causal control, this work traces how attention heads route identity, profile, and question information and how MLP blocks transform it into a character-conditioned response.

# 3 CoDePass: Role-Contrast Attribution over Complete Sequences

To isolate role-conditioned computation, we construct matched prompts that vary role information while preserving question semantics and the target response, then measure how each component reallocates support across the complete answer. This section introduces the analysis data and contrasts, defines the sequence-level estimator, and specifies the model–condition grid used in RQ1.

## 3.1 Analysis data and matched counterfactuals

We study characters from RoleBench's role-generalization split (Wang et al., 2024a) in two scenes. The *role-specific* scene asks about a character's experiences, relationships, or background, so answering requires character knowledge. The *general* scene asks ordinary questions that can be answered without this knowledge, while still requiring the model to respond as the assigned character. Each attribution cell contains the same quota of 200 matched examples, preventing characters with more available questions from dominating the ranking. The scene contrast separates profile-dependent response construction from persona modulation of a general answer.

For an intended role $r$, question $q$, and reference response $y$, the positive input $x^+$ contains the role instruction and profile. We construct three counterfactual inputs $x^-$: *no role* replaces the role instruction with a neutral assistant instruction and removes a leading direct role address when present, *wrong role* substitutes another identity and profile, and *no profile* retains the identity instruction but removes the profile. The reference response $y$ is held fixed within each pair. The contrasts therefore measure how role information reallocates support for the same answer, respectively probing role elicitation, persona distinction, and profile use. Appendices A.1, A.2, and A.3 report the data distributions, construction details, and a tokenized worked example.

## 3.2 CoDePass estimator

Role-conditioned computation develops throughout an answer, so single-position attribution can miss components that become important later. Contrastive Decomposed Forward Pass for Complete Sequences (CoDePass) extends the decomposed forward pass (Hong et al., 2025) to matched role contrasts over a complete reference response. Given $x^+$, $x^-$, and their shared response $y = (y_1, \ldots, y_{|y|})$, we append $y$ to both prompts and align target positions independently. At position $t$, the decomposed forward pass produces component $c$'s residual contribution $\mathbf{z}_{c,t}(x, y)$ to the prediction of $y_t$. CoDePass accumulates this contribution as

$$s_c(x, y) = \sum_{t=1}^{|y|} \langle \mathbf{z}_{c,t}(x, y), \mathbf{W}_U[y_t] \rangle \,. \tag{1}$$

Here, $\mathbf{W}_U$ is the model's output projection matrix, and $\mathbf{W}_U[y_t]$ is the unembedding vector for target token $y_t$. We contrast the two supports and normalize the absolute difference within each component family and example:

$$a_c(x^+, x^-, y) = \frac{\left| s_c(x^+, y) - s_c(x^-, y) \right|}{\sum_{c' \in \mathcal{C}} |s_{c'}(x^+, y) - s_{c'}(x^-, y)| + \epsilon}, \tag{2}$$

where $\mathcal{C}$ contains components of the same type and $\epsilon$ prevents division by zero. We average $a_c$ over examples and rank components by the mean, treating an attention head or a complete MLP block as one component. Heads and MLP blocks are normalized as separate families, and their spatial organization and concentration are analyzed within each family. Figure 8 and Algorithm 1 in

Appendices B.1 and B.2 combine full-response accumulation, matched role contrasts, and within-example normalization in one procedure. Appendix A.2 provides the complete prompt templates.

## 4 RQ1: Which Components Support Role Playing?

We begin with the first research question: *Which components support role playing?* We approach this through three systematic evaluations. First, we use CoDePass to isolate a shared role-playing backbone by comparing component attributions across models, characters, scenes, and role contrasts (Section 4.1). Next, progressive knockout experiments validate the structural necessity of these identified components (Section 4.2). Finally, cross-stage comparisons trace how this component backbone develops from pretraining to instruction tuning (Section 4.3). Appendix B.4 consolidates the model, data, diagnostic, optimization, and compute settings.

### 4.1 Detecting role-related components

To isolate the shared role-playing backbone from variable components, Figure 1 centers on a single reference condition: Qwen3-8B-Instruct (Yang et al., 2025) answering role-specific Jack Sparrow questions under a no-role contrast. Each subsequent panel perturbs exactly one factor to provide controlled comparisons against this baseline.

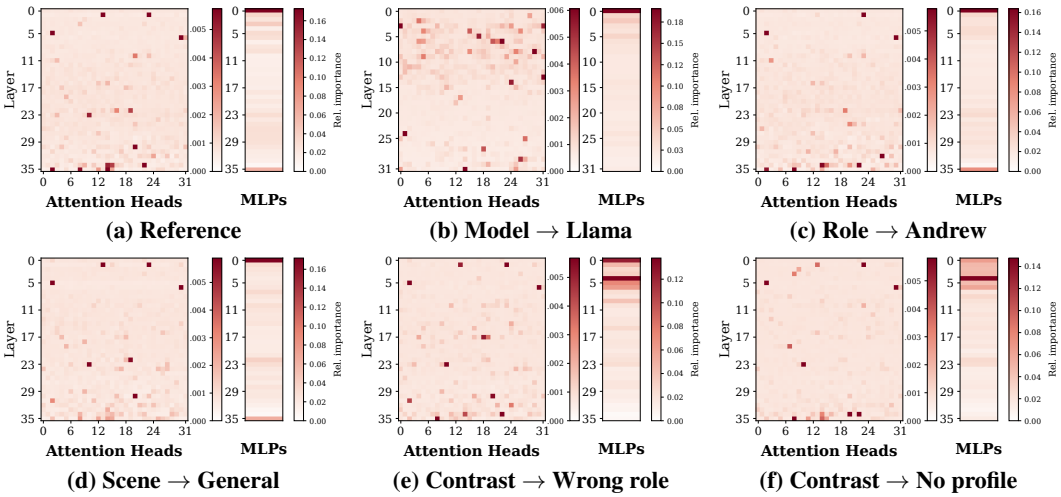



Figure 1: Importance of role-related components under controlled changes in model, character, scene, and role contrast. Squares denote attention heads and side strips denote MLP blocks; panel (a) is the reference and panels use independent scales.

**CoDePass identifies a compact, shared core of role-related attention heads.** Comparing characters (Figure 1a and c) reveals a consistent reliance on a highly sparse subset of early-layer heads, despite variations in later peaks. This significant overlap across different characters indicates that role-conditioned generation depends on a compact, reusable elicitation core complemented by character-specific components.

**This core-plus-variable organization transfers across architectures.** While Llama (Dubey et al., 2024) (Figure 1b) shifts the most prominent heads to deeper layers compared to the Qwen reference, both models maintain a consistent structural paradigm: recurring shared role-entry routes seamlessly coexist with character-sensitive persona pathways.

**Scene and contrast conditions disentangle role entry from persona discrimination.** Shifting to a general scene (Figure 1d) diminishes early shared peaks due to reduced profile reliance. Similarly, counterfactual contrasts reveal distinct mechanisms: omitting profiles (Figure 1f) preserves the elicitation core, whereas substituting the wrong role (Figure 1e) disperses attribution across later, discriminative heads. This indicates that profile use sharpens a reusable elicitation interface, while persona discrimination remains broadly distributed.

**MLP importance is broadly distributed across depth.** Unlike the compact attention routes, MLP role sensitivity (side strips, Figure 1a–f) spans the model's depth, maintaining stable contributions only near the boundaries with condition-dependent variations in between. This structural dichotomy—reusable head routes versus context-sensitive MLP allocations—motivates our two behavioral diagnostics in RQ2. Full layerwise MLP analyses, complete heatmaps, and extended summaries are detailed in Appendices B.3, C.1, C.2, and C.3.

### 4.2 Validating CoDePass through knockout

To validate the effectiveness of CoDePass in identifying components important for role playing, we progressively remove attention heads or MLP blocks in attribution-ranked order. We zero the selected components' outputs and compare the resulting ROUGE-L with removal of the same number of components in a fixed random order (Voita et al., 2019; Conmy et al., 2023; Lin, 2004). Appendix C.4 provides the complete intervention settings and generation outcomes.

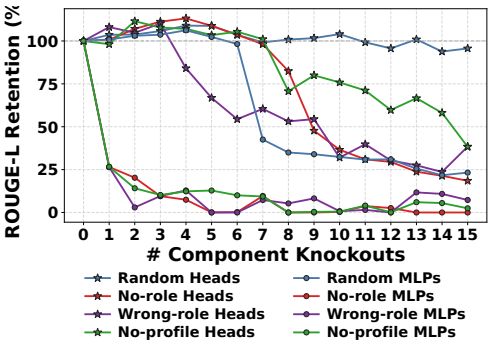



Figure 2: Role-playing performance under progressive component knockout, relative to the unmodified model. Each step removes 20 heads or one MLP block.

**Knockout results validate CoDePass attribution.** As shown in Figure 2, ablating CoDePass-ranked components causes significantly steeper performance degradation than random removal. Removing top-ranked heads leads to a severe collapse in ROUGE-L scores, while ablating even the first few ranked MLPs disrupts performance immediately. This higher sensitivity to MLP removal highlights their critical structural role. Overall, these targeted degradation patterns confirm that CoDePass effectively isolates components essential for role-playing.

### 4.3 Consistency of role-related components across training stages

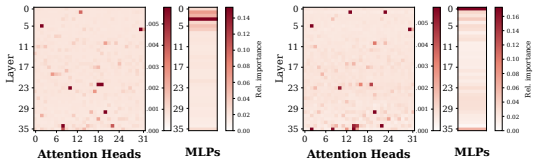



Figure 3: Importance of role-related components in Qwen3-8B Base and Instruct for Jack Sparrow. Panels use independent scales.

To investigate the developmental trajectory of role-related components, we compare Qwen3-8B Base and Instruct under strictly controlled conditions.

**Instruction tuning reorganizes attention routes while preserving the MLP backbone.** As illustrated in Figure 3 and supported by Table 1, the transition from Base to Instruct induces a statistically significant distributional shift in attention heads, whereas MLP allocations remain relatively stable. Although a core subset of top-ranked components is shared across both stages, instruction tuning substantially rewires the attention pathways connecting role information to answer preparation, leaving the pretrained MLP transformation backbone largely intact. Complete component ranks are reported in Appendix C.5.1.

Table 1: Statistical comparison of component attribution between Qwen3-8B Base and Instruct.

| Component | K–S Statistic | K–S Test $p$-value | Changed Components | Maximum Attribution Change |
|---|---|---|---|---|
| Attention heads | .186 | $\sim 10^{-17}$ | 14 of 32 | L6H30 ($-6.67$) |
| MLP blocks | .306 | .069 | 3 of 8 | L3 ($-4.29$) |

**Role-playing emerges through a two-stage developmental process.** Comparisons with a randomly initialized baseline further elucidate these dynamics. Pretraining aggressively concentrates attention mass into a sparse, role-sensitive scaffold, completely overwriting the random initialization. Subsequently, instruction tuning refines this scaffold, redistributing attention across early role entry, middle-layer coordination, and final answer preparation. These results reveal a clear two-stage evolution: pretrain-

ing establishes the fundamental routing architecture, and instruction tuning aligns it into a coordinated role-playing system. Appendix C.5.2 details the complete random–Base–Instruct comparison. Building on this structural understanding, RQ2 next asks what specific information these components route and transform.

# 5 RQ2: What Patterns Do Role-Related Components Exhibit?

RQ1 revealed a sparse attention-routing scaffold and role sensitivity distributed across MLP depth. We now address the second question: *What functional patterns do these role-related components exhibit?* We use two complementary diagnostics. Value-weighted token-interaction maps identify the prompt information retrieved by localized heads (Section 5.1), while output-embedding projections trace how MLPs reshape role-conditioned states across depth (Section 5.2).

## 5.1 Analysis of Attention Heads

Attention probabilities reveal where a head attends but not the strength of the value retrieved from that position (Kobayashi et al., 2020). For head $h$ in layer $\ell$, let $A_{qk}^{\ell,h}$ be the causally masked attention probability from query $q$ to key $k$, and let $\mathbf{v}_k^{\ell,h}$ be the corresponding transformed value vector. We compute the full interaction map

$$O_{qk}^{\ell,h} = A_{qk}^{\ell,h} \left\| \mathbf{v}_k^{\ell,h} \right\|_2, \tag{3}$$

using the grouped-query value head associated with query head $h$ (Ainslie et al., 2023), where $N$ is the sequence length. Unlike a regional total, $\mathbf{O}^{\ell,h} \in \mathbb{R}^{N \times N}$ exposes both vertical token-selection patterns and query–key positional structure. Appendix D.1 specifies the diagnostic protocol, sampling, and frozen-class criteria.

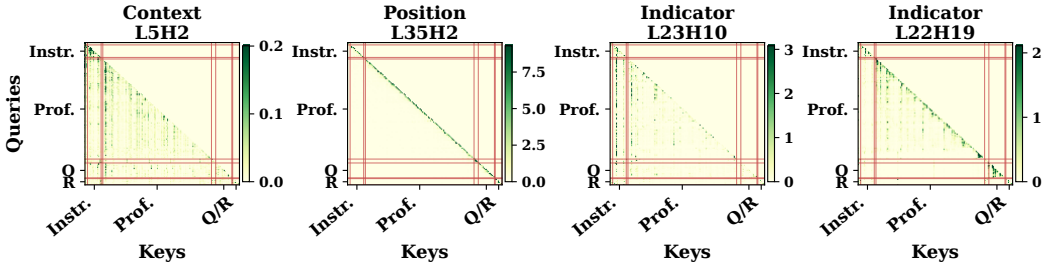



Figure 4: Value-weighted attention patterns of role-related heads. Q denotes question tokens; R denotes the final eight scaffold tokens at the response boundary.

**Attribution-ranked heads exhibit three specialized functions.** Complete interaction maps reveal three repeated behaviors. Vertical profile and question bands identify *role-context heads* that repeatedly retrieve character information and query content. Identity-token columns identify *role-indicator heads* that select persona cues. Local diagonals identify *positional heads* that carry neighboring-token information and anchor the response boundary (Figure 4). Each class concentrates most strongly in the region that defines it, dividing role processing into context retrieval, identity-cue selection, and answer-boundary timing.

**The functions persist while their inputs track the scene and persona.** We define the three classes once in the discovery condition and transfer the assignments unchanged to the remaining characters. On general questions, context heads exchange profile mass for question mass, indicator heads reduce identity concentration, and positional heads preserve their response-boundary pattern (Figure 5). Across held-out roles and scenes, context and positional functions remain stable, while identity selection transfers only partly. Content and timing follow the shared prompt schema; identity control must be rebound for each persona. Role routing therefore combines reusable operations with a character-dependent control signal, explaining why RQ1 finds recurrent elicitation/profile routes but variable wrong-role distinctions. Appendix D.2 reports the discovery-role scene comparison and complete held-out transfer results.

### 5.2 Analysis of MLPs

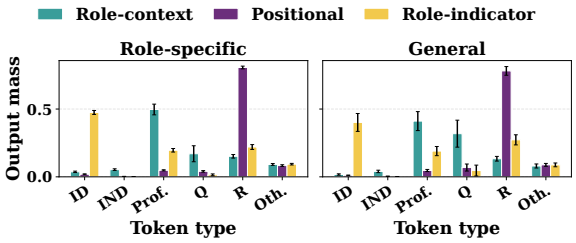
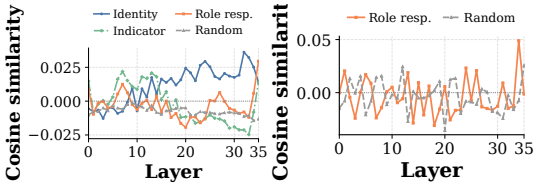



Figure 5: Regional attention-value mass across head classes. ID and IND mark identity and role-indicator tokens; Prof., Q, R, and Oth. mark profile, question, response-boundary, and other tokens.

Heads reveal what information is routed from the prompt; MLP probes ask how the resulting residual state changes across depth. Indicator and identity directions test whether an explicit role instruction develops into an identity-conditioned answer state, while the response direction locates answer preparation. Following vocabulary- and embedding-space analyses of Transformer parameters (Geva et al., 2022; Dar et al., 2023), let $\mathbf{u}_T$ be the mean output embedding for token set $T$, let $\mathrm{MLP}_{in}$ and $\mathrm{MLP}_{out}$ denote the residual states immediately before and after the block, and define $\mathbf{\Delta}_{\mathrm{MLP}} = \mathrm{MLP}_{out} - \mathrm{MLP}_{in}$. At the response-preparation position, $\langle \mathrm{MLP}_{in}, T \rangle$ is the cosine similarity between the pre-MLP residual and $\mathbf{u}_T$, while

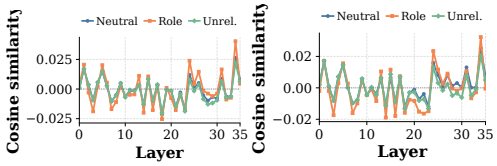
$$\langle \mathrm{MLP}_{out} - \mathrm{MLP}_{in}, T \rangle = \cos(\mathbf{\Delta}_{\mathrm{MLP}}, \mathbf{u}_T) \tag{4}$$

isolates the MLP update. Role-related directions are paired with equal-length random-token controls. Appendix D.3 gives the complete probe and control construction.




(a) Reception (b) Generation

Figure 6: MLP reception and generation of role-related information. Reception probes the pre-MLP residual; generation probes the MLP update.




(a) Jack Sparrow (b) Prof. Dorr

Figure 7: MLP alignment with neutral and persona-rendered response directions across layers.

**MLPs organize role information in three depth regimes.** Figure 6a reveals a clear depthwise progression: early residuals align with explicit role indicators, middle layers shift toward identity alignment, and final layers peak in response preparation. This ordered transition—absent in random controls—mirrors the routing sequence of attention heads, jointly supporting a staged evolution from role cues to identity-conditioned answers.

**Role-response directions are refined repeatedly across the stack.** As shown in Figure 6b, successive MLPs alternately add and subtract response-aligned components, culminating in a dominant write to the final layer. Combined with the knockout sensitivity observed in RQ1, this oscillating pattern characterizes response formation as a distributed, stack-wide refinement process rather than a localized operation.

**Characters share a semantic trajectory but vary in persona-rendering timing.** Figure 7 identifies two distinct trajectory families across characters: some integrate persona cues immediately, while others exhibit a brief neutral semantic lead before diverging. Crucially, all characters eventually converge on a shared layerwise geometry for semantic construction, with persona rendering acting as a character-dependent timing offset (Appendix D.4).

**Reusable routing functions transfer across architectures.** Cross-architecture evaluations on Llama-3.1-8B-Instruct (Appendices D.2 and D.4) reveal that context and positional functions are reliably maintained. This indicates that content retrieval and timing routes are architecturally robust, whereas identity selection and late persona-sensitive transformations remain model-specific.

**Role processing relies on a coherent cue-to-identity-to-answer pipeline.** In summary, early attention heads route profile data, middle MLPs construct identity alignment, and late MLPs refine the final response. Because character variations predominantly alter attention-driven identity selection

rather than the shared MLP semantic trajectory, we hypothesize that targeted attention updates will drive role more efficiently than expanding MLP capacity. RQ3 tests this hypothesis.

# 6 RQ3: Can Component-Guided Tuning Improve Role Playing?

The first two questions identified a recurring role-elicitation scaffold, separable attention-routing functions, and a distributed cue-to-identity-to-answer transformation across MLP depth. We now ask the final question: *Can tuning these components improve role playing?* We select query-head groups from the shared role-specific/general no-role ranking because RQ1 found this contrast to expose recurrent cross-character elicitation routes, and compare CoDePass-guided tuning with full-parameter and capacity-matched random tuning on unseen characters.

## 6.1 Experimental Setup and Results

**Experimental setup.** We examine three approaches on Qwen3-8B-Instruct: full-parameter fine-tuning (Full SFT), CoDePass-guided fine-tuning (Targeted SFT), and capacity-matched random-component fine-tuning (Random SFT). The shared ranking averages CoDePass matrices equally over 75 characters and both scenes; the same characters provide SFT data, and six disjoint characters form the evaluation set. Targeted SFT updates the Q/K/V/O slices of the top 64 query-head groups, while Random SFT matches their layerwise query-head counts, grouped-query K/V groups, and parameter count. We evaluate RoleBench in both scenes with ROUGE-L and SacreBLEU (Post, 2018), and general capabilities on MMLU, ARC-Challenge, and HellaSwag (Hendrycks et al., 2021; Clark et al., 2018; Zellers et al., 2019). Scores use a 0–100 scale, and these results report means and sample standard deviations over three seeds. Appendix E.1 provides the complete protocol.

Table 2: Overall evaluation on role-playing and generic tasks with Qwen3-8B-Instruct. ± denotes the sample standard deviation over three training seeds.

| Models | Train Time (ms/sample)↓ | Tuned Parameters↓ | Role-Playing Tasks | | | | Generic Tasks | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Role-Specific | | General Scene | | MMLU | ARC | HellaSwag |
| | | | ROUGE-L↑ | SacreBLEU↑ | ROUGE-L↑ | SacreBLEU↑ | Acc.↑ | Acc.↑ | Acc.↑ |
| Qwen3-8B-Instruct | – | – | 18.09 | 3.30 | 11.84 | 2.68 | 71.94 | 63.31 | 55.87 |
| + Full SFT | 50.3 | 8.19B (100%) | 22.42±0.11 | 4.28±0.09 | 35.91±0.20 | 11.36±0.71 | 73.85±0.21 | 65.90±0.18 | 73.85±0.09 |
| + Random SFT | 32.9 | **124.8M (1.52%)** | 21.31±0.10 | 3.95±0.30 | 34.07±0.74 | 10.08±0.43 | 73.32±0.25 | 64.61±0.10 | 73.46±0.11 |
| **+ Targeted SFT** | **32.3** | **124.8M (1.52%)** | **23.39±0.11** | **4.32±0.06** | **39.08±0.12** | **13.34±0.19** | **74.21±0.12** | **65.93±0.32** | **75.16±0.31** |

**Experimental results.** Table 2 shows three benefits of Targeted SFT: (1) **Improved Role-Playing Performance**: Targeted SFT raises role-specific ROUGE-L by 9.8% over Random SFT and 4.3% over Full SFT. On general-scene questions, the corresponding gains are 14.7% and 8.8%, with higher SacreBLEU as well. (2) **Preservation of General Capabilities**: Targeted SFT improves all three generic-task scores over the original model and achieves the highest MMLU, ARC and HellaSwag scores. (3) **Enhanced Training Efficiency**: Targeted SFT updates 1.52% of the parameters and reduces training time by 35.9% relative to Full SFT. Additional model results appear in Appendix E.2.

## 6.2 Ablation Study of Trainable Components

Table 3: Ablative experiments on attention heads.

| Attention Heads | Parameters | Role ROUGE-L |
|---|---|---|
| QKVO-8 | 16.8M (0.20%) | 21.25±0.28 |
| QKVO-16 | 32.5M (0.40%) | 21.87±0.20 |
| QKVO-32 | 65.0M (0.79%) | 22.95±0.21 |
| Random QKVO-64 | 124.8M (1.52%) | 21.31±0.10 |
| Full SFT | 8.19B (100.00%) | 22.42±0.11 |
| **QKVO-64** | 124.8M (1.52%) | **23.39±0.11** |

Table 4: Ablative experiments on MLPs.

| MLP Configuration | Parameters | Role ROUGE-L |
|---|---|---|
| QKVO-64 | 124.8M (1.52%) | 23.39±0.11 |
| + Top-1 MLP | 275.8M (3.37%) | **23.70±0.23** |
| + Top-3 MLPs | 577.8M (7.05%) | **23.70±0.30** |
| + Top-5 MLPs | 879.8M (10.74%) | 22.70±0.40 |
| + Random-3 MLPs | 577.8M (7.05%) | 23.62±0.36 |
| Full SFT | 8.19B (100.00%) | 22.42±0.11 |

To test how targeted fine-tuning depends on trainable capacity, we sweep 8–160 ranked QKVO groups and add top-ranked MLP blocks (full results in Appendix E.4). Table 3 shows that performance improves with some fluctuations as ranked head coverage expands to 160 heads, surpassing matched Random and Full SFT. Expanding to 96–160 heads brings little further improvement. This

suggests that attention-head effects are sparse at model scale but depend on coverage within the selected set. In contrast, adding MLP blocks improves performance only when a few are added (Table 4). Together with RQ2, these results show that the existing MLP pathway supports role-playing through attention updates, with selective MLP refinement providing a smaller benefit.

### 6.3 Further Analysis of Transfer and Answer Content

We examine whether selected components transfer beyond discovery characters and whether tuning more ranked heads retains relevant content with fewer words. Appendices E.3 and E.4 further examine random component selection and the relationship between head coverage, answer content, and response length.

Table 5: Cross-role transfer of Targeted SFT on ranking and held-out characters. Brackets give crossed 95% confidence intervals.

| Evaluation Characters | vs. Random SFT | vs. Full SFT |
|---|---|---|
| Ranking Characters | $+3.14$ $[2.31, 3.98]$ | $+2.09$ $[1.39, 2.82]$ |
| Held-out Characters | $+2.08$ $[1.63, 2.62]$ | $+0.97$ $[0.48, 1.36]$ |

Table 6: Answer precision, recall, and length after fine-tuning.

| Method | Precision | Recall | Words | ROUGE-L |
|---|---|---|---|---|
| Five-head Q/O | 30.81±0.65 | **17.63±0.09** | 52.9±3.3 | 21.05±0.28 |
| Targeted QKVO-64 | **39.49±0.46** | 17.22±0.10 | 28.9±0.3 | **23.39±0.11** |
| Full SFT | 37.70±0.57 | 16.69±0.16 | 31.2±1.2 | 22.42±0.11 |

**The selected routes transfer to unseen characters.** Targeted SFT outperforms Random and Full SFT on development questions for ranking characters and test questions for held-out characters (Table 5). Gains persist with disjoint ranking and evaluation characters, indicating a reusable role-routing interface.

**Broader head coverage produces shorter answers that retain relevant content.** The five-head Q/O intervention recovers reference content but produces longer, less precise answers. QKVO-64 raises precision from 30.81 to 39.49 with similar recall and shortens responses from 52.9 to 28.9 words (Table 6). Updating K/V within the same five-head set improves general-scene ROUGE-L but lowers role-specific ROUGE-L. Broader ranked coverage therefore improves role-specific content selection beyond expanding the updated projections alone.

### 6.4 Supplementary Analyses

To connect the routing functions identified in RQ2 to tuning outcomes and to characterize generation quality, Appendices E.5 and E.6 detail functional-class controls and qualitative case studies. Functional interventions reveal that updating role-indicator heads provides the most effective single-class role control. However, combining functional classes yields non-additive gains, confirming that the data-driven Top-64 selection more efficiently captures necessary routing synergies. Qualitatively, Targeted SFT significantly reduces verbosity, resolves severe length mismatches, and better aligns answers with character-specific causal reasoning and beliefs. Despite these gains, remaining errors—such as the dilution of idiosyncratic character voices and the unintentional smoothing of antagonistic personas into generic, egalitarian viewpoints—highlight key challenges for future role-playing alignment.

## 7 Conclusion

This study systematically explores the role-playing mechanisms of LLMs by progressively addressing three research questions. We first identify crucial components using CoDePass, revealing that role adaptation relies on a sparse subset of attention heads and distributed MLPs. These components exhibit a clear division of labor: localized attention heads retrieve and route identity cues into the response pathway, while MLPs progressively transform these signals into role-conditioned representations. Guided by these insights, we find that targeted fine-tuning of as little as 2% of the parameters surpasses both capacity-matched random and full-parameter tuning on held-out characters while preserving general capabilities. These results provide a unified account of how role information is selected, transformed, and controlled during generation.

## AI USE STATEMENT

Generative AI tools are used to polish the manuscript's language, refine its structural presentation, and assist in writing code for the experimental implementation. Following the use of these tools, the human authors conduct a complete and careful review of all generated text and code. The authors assume full responsibility for the accuracy, originality, and integrity of the final manuscript and its associated codebase.

## REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code in an anonymous repository at `https://anonymous.4open.science/r/CoDePass-0E45`.

## REFERENCES

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325. Association for Computational Linguistics, 2024.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901. Association for Computational Linguistics, 2023.

Rumi Allbert, James K. Wiles, and Vlad Grankovsky. Identifying and manipulating personality traits in LLMs through activation engineering. *arXiv preprint arXiv:2412.10427*, 2025.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126. Association for Computational Linguistics, 2024.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Celia Cintas, Miriam Rateike, Erik Miehling, Elizabeth Daly, and Skyler Speakman. Localizing persona representations in LLMs. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):630–642, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170. Association for Computational Linguistics, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. 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)*, pp. 6218–6240. Association for Computational Linguistics, 2025.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Xiangyu Hong, Che Jiang, Kai Tian, Biqing Qi, Youbang Sun, Ning Ding, and Bowen Zhou. DePass: Unified feature attributing by simple decomposed forward pass. In *Advances in Neural Information Processing Systems 38*, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7057–7075. Association for Computational Linguistics, 2020.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. RoleMRC: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21008–21030. Association for Computational Linguistics, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Ansh Poonia and Maeghal Jain. Dissecting persona-driven reasoning in language models via activation patching. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24553–24566. Association for Computational Linguistics, 2025.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191. Association for Computational Linguistics, 2018.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14467–14479. Association for Computational Linguistics, 2024.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187. Association for Computational Linguistics, 2023.

Yihong Tang, Kehai Chen, Muyun Yang, Zheng-Yu Niu, Jing Li, Tiejun Zhao, and Min Zhang. Thinking in character: Advancing role-playing agents with role-aware reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 130522–130556, 2025.

Yihong Tang, Kehai Chen, Xuefeng Bai, Benyou Wang, Zeming Liu, Haifeng Wang, and Min Zhang. Character-R1: Enhancing role-aware reasoning in role-playing agents via RLVR. *arXiv preprint arXiv:2601.04611*, 2026a.

Yihong Tang, Kehai Chen, Liang Yue, Benyou Wang, and Min Zhang. CRPO: Character-centric group relative policy optimization for role-aware reasoning in role-playing agents. In *Forty-third International Conference on Machine Learning*, 2026b.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850. Association for Computational Linguistics, 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Miles Wang, Tom Dupre la Tour, Olivia Watkins, Aleksandar Makelov, Ryan Andrew Chi, Samuel Miserendino, Jeffrey George Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Daniel P Mossing. Persona features control emergent misalignment. In *The Fourteenth International Conference on Learning Representations*, 2026.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777. Association for Computational Linguistics, 2024a.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873. Association for Computational Linguistics, 2024b.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64822–64858. PMLR, 2025a.

Yongjie Wang, Jonathan Leung, and Zhiqi Shen. RoleRAG: Enhancing LLM role-playing via graph guided retrieval. *arXiv preprint arXiv:2505.18541*, 2025b.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12540–12557. Association for Computational Linguistics, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-Ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. Interpreting and improving large language models in arithmetic calculation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59932–59950. PMLR, 2024.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1457–1476. Association for Computational Linguistics, 2024.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26101–26110, 2025.

# A Role-Playing Data and Prompt Templates

## A.1 Dataset construction and statistics

**Data for RQ1 and RQ2.** We construct the analysis dataset from the RoleBench role-generalization split (Wang et al., 2024a). It contains six fictional characters and two question scenes. Role-specific questions ask about a character's experiences, relationships, or identity, whereas general questions contain character-independent instructions that can still be answered in the assigned persona. Table 7 reports the complete distribution. The analysis set contains 4,549 questions and between one and five reference responses per question.

Table 7: RoleBench analysis-set statistics. Reference columns give means per question.

| Character | Role-Specific | General | References/Role | References/General |
|---|---|---|---|---|
| Jack Sparrow | 286 | 500 | 1.01 | 4.57 |
| Professor G.H. Dorr | 201 | 500 | 1.00 | 4.71 |
| Andrew Detmer | 264 | 500 | 1.00 | 3.79 |
| Coriolanus | 302 | 500 | 1.01 | 4.20 |
| Tom Ripley | 287 | 500 | 1.01 | 3.79 |
| John Keating | 209 | 500 | 1.00 | 4.07 |
| Total | 1,549 | 3,000 | – | – |

Role-specific availability varies across characters, whereas the general scene contains 500 questions for every role and more references per question. We therefore use the same 200-example quota in every attribution cell and aggregate condition statistics at the role level. This prevents characters with more available questions or references from dominating the cross-role component rankings.

**Data for RQ3.** The main component-guided fine-tuning experiment uses the 75 characters shared by the role-specific and general training splits. These characters provide both the CoDePass ranking data and the SFT corpus, aligning component discovery with the supervised task. The six analysis characters form a disjoint evaluation set for testing unseen characters. The training and attribution splits balance both scenes, as shown in Table 8.

Table 8: Role-playing SFT corpus statistics.

| Split | Examples | Characters | Role-Specific | General |
|---|---|---|---|---|
| Training | 18,205 | 75 | 9,107 | 9,098 |
| Development (Attribution) | 994 | 75 | 492 | 502 |

## A.2 Role-contrast templates

**Prompt templates.** For role $r$, profile $p_r$, question $q$, substituted role $r'$, and target response $y$, Table 9 gives the system-instruction templates. Each instruction-tuned input consists of the selected system instruction, the user message $q$, and the assistant-generation boundary; $y$ is then appended as the shared attribution target. The question and target response are held fixed across the positive, wrong-role, and no-profile conditions. For role-specific no-role prompts, a leading direct address to $r$ is removed from the question; general questions remain unchanged. Six characters, two scenes, and three counterfactuals yield 36 matched-pair groups for each model.

Table 9: System prompts for CoDePass role contrasts.

| Condition | System Prompt |
|---|---|
| Positive | `You are` $r$`. Stay in character and answer as` $r$`.` `Profile:` $p_r$ |
| No role | `You are a helpful assistant.` |
| Wrong role | `You are` $r'$`. Stay in character and answer as` $r'$`.` `Profile:` $p_{r'}$ |
| No profile | `You are` $r$`. Stay in character and answer as` $r$`.` |

The three contrasts isolate different role-playing signals. Positive versus no role measures role elicitation, positive versus wrong role measures persona distinction, and positive versus no profile measures the use of character background. In every contrast, the same reference response $y$ defines the sequence-level attribution target.

**Counterfactual construction.** No-role counterfactuals replace the system instruction with a neutral assistant instruction and remove a leading character name from role-specific questions. Wrong-role counterfactuals substitute the paired character's complete identity and profile according to Table 10. No-profile counterfactuals preserve the identity instruction and remove the profile.

Table 10: Persona substitutions used in wrong-role counterfactuals.

| Intended Role | Counterfactual Role |
|---|---|
| Jack Sparrow | Tom Ripley |
| Professor G.H. Dorr | John Keating |
| Andrew Detmer | Coriolanus |
| Coriolanus | Professor G.H. Dorr |
| Tom Ripley | Jack Sparrow |
| John Keating | Andrew Detmer |

Every character appears once as an intended role and once as a counterfactual role. The balanced substitution scheme prevents one especially similar or dissimilar distractor persona from dominating the wrong-role comparison.

### A.3 Token types and worked example

Consider the role-specific question "Jack Sparrow, how did you manage to escape from the casket?" The positive prompt assigns Jack Sparrow and provides his complete profile. Its reference begins, "Ah, well, you see, I always have a few tricks up my sleeve, or in this case, up my coffin." Table 11 divides the rendered prompt into the regions used by the attention and MLP analyses.

Table 11: Token counts by input region in Qwen3-8B-Instruct and Llama-3.1-8B-Instruct.

| Region | Example Content | Qwen | Llama |
|---|---|---|---|
| Role instruction | `You are Jack Sparrow. Stay in character...` | 16 | 16 |
| Profile | Character background and traits | 128 | 128 |
| Question | `Jack Sparrow, how did you manage...` | 15 | 15 |
| Chat scaffold | Message boundaries and assistant header | 19 | 37 |
| Total prompt | | 178 | 196 |

Table 12 shows the actual token decomposition around the response boundary. The symbol ␣ marks a tokenizer word boundary, and NL denotes a line-break token.

Table 12: Tokens at the question-to-response boundary in Qwen3-8B-Instruct and Llama-3.1-8B-Instruct.

| Token Type | Decoded Token Sequence |
|---|---|
| Qwen question/boundary | ␣c, asket, ?, <\|im_end\|>, NL, <\|im_start\|>, assistant, NL, <think>, NL×2, </think>, NL×2 |
| Llama question/boundary | ␣c, asket, ?, <\|eot_id\|>, <\|start_header_id\|>, assistant, <\|end_header_id\|>, NL×2 |
| Target response | Ah, ,, ␣well, ,, ␣you, ␣see, ,, ␣I, ␣always, ␣have, ␣a, ␣few, ␣tricks, ... |

The response-preparation query is the final scaffold token, at position 177 for Qwen3-8B-Instruct and 195 for Llama-3.1-8B-Instruct. Qwen's boundary contains the assistant marker followed by an empty thinking block; Llama's boundary contains the assistant header. CoDePass begins with the first reference-response token and accumulates contributions over the complete response.

# B Experimental Details of CoDePass

## B.1 Overview of CoDePass

CoDePass compares two decomposed forward passes with a shared target response. Figure 8 shows how a head writes to the residual stream and how its contribution propagates through subsequent attention and MLP blocks to the output. The same decomposition can start from a complete MLP update.

## B.2 Attribution procedure

For instruction-tuned models, the chat template receives system and user messages followed by the assistant generation marker. Base-model prompts concatenate the system instruction, question, and answer marker. Positive and counterfactual prompts are tokenized independently, and their target positions are aligned to the shared response. Equation 1 accumulates each component's contribution over the response; Equation 2 then normalizes the contrast separately for attention heads and MLP blocks.

**Algorithm 1: CoDePass**

**Input:** role $r$, question $q$, response $y$, contrast $m$, model $f$, component family $\mathcal{C}$.
**Output:** normalized importance $a_c$ for every $c \in \mathcal{C}$.

1. Construct matched prompts $x^+ = \mathrm{Prompt}(r, q)$ and $x^- = \mathrm{Counterfactual}(x^+, m)$.
2. Append the same response $y$ to both prompts and align target positions independently.
3. Run decomposed forward propagation to obtain $\mathbf{z}^+_{c,t}$ and $\mathbf{z}^-_{c,t}$.
4. Accumulate sequence support with Equation 1.
5. Normalize the support difference with Equation 2.
6. Average $a_c$ over examples and rank heads or MLP blocks.

## B.3 MLP granularity

MLP attribution treats each feed-forward block as one component and produces one importance score per layer. The behavioral analysis subsequently separates the residual state entering each MLP from the update written by that block, allowing the layerwise transformation process to be traced across depth.

## B.4 Experimental setup details

Table 13 consolidates the model, data, behavioral-analysis, optimization, evaluation, decoding, and compute settings shared across the three research questions. The controlled RQ1 and RQ2 diagnostics retain the same six-character inputs and generation protocol. RQ3 expands component discovery

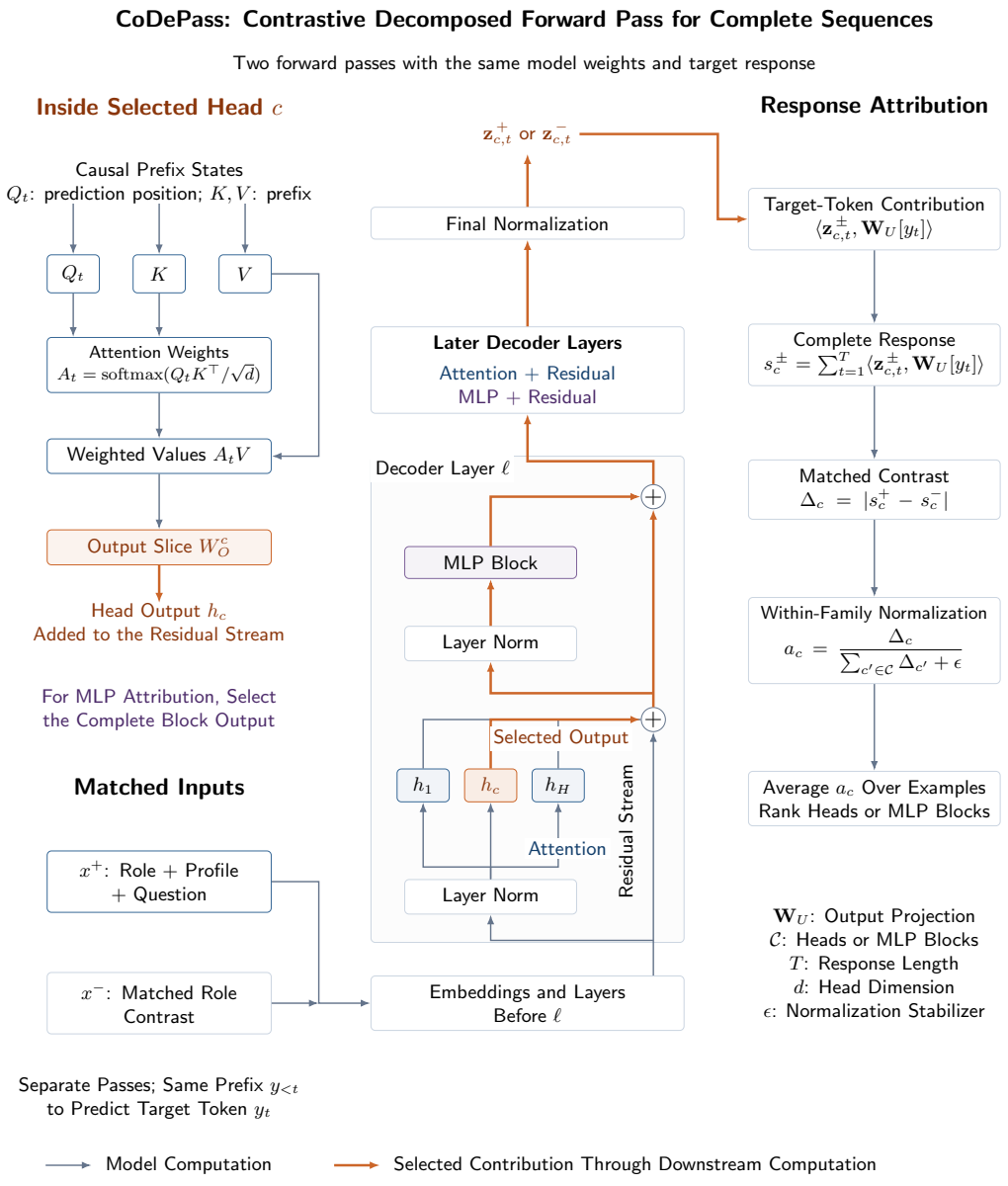



Figure 8: CoDePass inside a decoder-only Transformer. Orange paths track a selected component's contribution through downstream layers. Both passes use the same weights and reference response; the inset expands one attention head.

and SFT to 75 shared training characters, then evaluates both scenes on the six characters excluded from training.

## C Component Localization and Intervention Results

This appendix extends RQ1 with the complete attribution grid, cross-character and scene summaries, progressive-knockout outcomes, and the Base-to-Instruct comparison.

Table 13: Experimental settings for the three research questions.

| Item | Setting |
|---|---|
| Models | Qwen3-8B Base and Instruct (Yang et al., 2025), and Llama-3.1-8B-Instruct; RQ3 additionally uses Llama-3.1-8B-Base (Dubey et al., 2024); bfloat16 forward computation |
| Attribution Data | RQ1–RQ2: six characters, two scenes, and three matched contrasts; RQ3: 75 training characters, two scenes, and the no-role contrast |
| Behavioral Analysis | Response-preparation query; value-weighted attention; MLP residual updates at block granularity |
| Fine-tuning | 75 characters in both scenes; two epochs for Qwen and one for Llama; global batch size 64; AdamW; cosine decay; 2% warmup; maximum sequence length 1,024 |
| Evaluation | First-reference ROUGE-L and SacreBLEU; MMLU 5-shot, ARC-Challenge 25-shot, and HellaSwag 10-shot |
| Decoding | Greedy generation with thinking disabled and responses constrained to at most 120 words |
| Compute | Eight NVIDIA H100 80GB GPUs |

### C.1 Cross-role overlap and scene effects

**Cross-role overlap.** To distinguish shared role-playing components from character-specific ones, we compare the leading component sets across characters under each role contrast. Table 14 quantifies how often the most important components recur across the six roles. Wrong-role comparisons have the lowest head overlap for all three models in the role-specific scene. No-role and no-profile perturb whether a shared role mode or its character information is available, so common elicitation and retrieval routes recur across personas. Wrong-role preserves a complete role prompt on both sides and subtracts away more of that shared computation, exposing character-dependent identity discrimination. This explains the more distributed persona-boundary pattern in Figure 1.

Table 14: Mean pairwise Jaccard overlap of top-ranked components across characters on role-specific questions.

| | Top-20 Attention Heads | | | Top-5 MLP Blocks | | |
|---|---|---|---|---|---|---|
| Model | No Role | No Profile | Wrong Role | No Role | No Profile | Wrong Role |
| Qwen Instruct | .495 | .364 | .299 | .540 | .272 | .253 |
| Qwen Base | .447 | .248 | .229 | .311 | .278 | .280 |
| Llama Instruct | .274 | .398 | .127 | .488 | .328 | .206 |

**Scene effects.** To determine whether the need for character knowledge changes attribution concentration, we compare the leading head mass between role-specific and general questions. For Qwen-Instruct, the mean top-10 head mass is 15.81% for role-specific/no-role conditions and 9.48% for general/no-role conditions. No-profile shows a similar change (15.50% versus 10.03%), while wrong-role remains nearly constant (7.88% versus 8.07%). Role-specific questions therefore sharpen shared role elicitation and profile retrieval, whereas the persona-boundary signal remains distributed in both scenes. This interaction shows that profile dependence changes the concentration of shared routing, while identity discrimination remains necessary even when the answer content is character-independent.

### C.2 Top Qwen-Instruct candidates by contrast

To identify which components respond to different kinds of role information, we compare the highest-ranked heads and MLP blocks across the three contrasts. Table 15 lists the top five components after averaging the six roles in the role-specific scene. The same three early heads lead every contrast, but their joint mass falls sharply for wrong-role comparisons. Later heads such as L34H22 and L35H8 enter the leading set under no-profile or wrong-role comparisons. MLP rankings change more substantially: only L0 remains in the top five for every contrast, while the remaining blocks reorganize with the role information being removed or replaced. The head core therefore provides

shared entry routes, whereas the distributed MLP pathway adapts more strongly to the contrast semantics.

Table 15: Top Qwen3-8B-Instruct components under each role contrast.

| Contrast | Attention Heads (Attribution %) | MLP Blocks (Attribution %) |
|---|---|---|
| No role | L5H2 (4.959), L6H30 (4.332), L1H23 (4.286), L1H13 (.532), L35H8 (.340) | L0 (6.811), L35 (5.911), L1 (3.485), L3 (3.240), L5 (3.237) |
| No profile | L6H30 (4.703), L5H2 (4.669), L1H23 (3.413), L1H13 (.495), L34H22 (.487) | L0 (5.204), L4 (3.375), L8 (3.159), L7 (3.159), L22 (3.084) |
| Wrong role | L1H23 (2.173), L5H2 (1.818), L6H30 (1.778), L34H22 (.410), L35H8 (.318) | L0 (5.358), L3 (3.360), L4 (3.339), L2 (3.226), L5 (2.986) |

### C.3 Complete attribution heatmaps

To examine how consistently the localization patterns hold across characters, scenes, and models, we visualize the complete attribution grid below.

**Qwen3-8B-Instruct separates shared role entry from persona distinction.** Figures 9–14 show recurring early peaks across characters, especially for no-role and no-profile contrasts. Role-specific questions concentrate these two contrasts because their answers require role elicitation and profile information. Wrong-role attribution spreads over a broader late-layer set: both prompts already elicit role behavior, so the contrast removes their shared format and emphasizes identity discrimination. On general questions, no-role and no-profile attribution becomes less concentrated because generic semantics can form the answer, while persona information acts as a lighter modulation.

**Qwen3-8B Base contains the scaffold before instruction tuning.** Figures 15–20 show prominent early and late routes in the Base model, including heads that remain highly ranked in the Instruct system. The raw aggregate ranking places the response-interface head L1H23 at 452 in Base and 2 in Instruct, while L5H2 and L6H30 remain among the leading heads. The Base-to-Instruct system transition therefore preserves a role-sensitive scaffold while rebalancing it toward response-interface coordination.

**Llama preserves the organization but relocates its components.** Figures 21–26 are less concentrated at the head level and use different coordinates from Qwen. Within role-specific questions, no-profile heads recur most strongly across characters and wrong-role heads least strongly, preserving the separation between reusable profile/elicitation operations and persona-dependent identity boundaries. MLP mass remains distributed across depth. The cross-architecture regularity lies in this organization: both models combine recurring role routes with condition-sensitive components, with model-specific locations and concentrations.

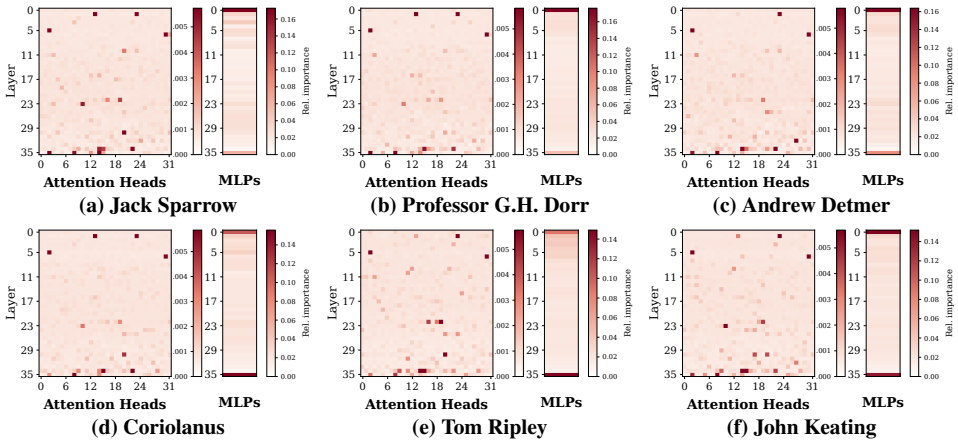



Figure 9: Importance of role-related components in Qwen3-8B-Instruct on role-specific questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

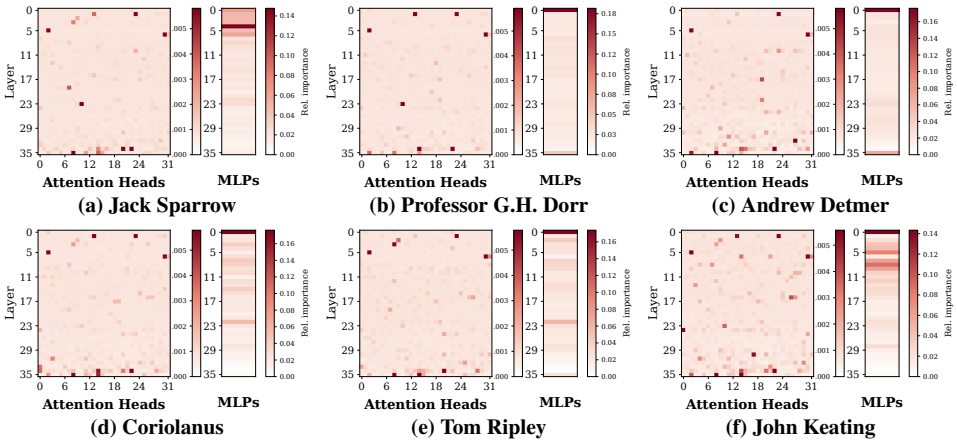



Figure 10: Importance of role-related components in Qwen3-8B-Instruct on role-specific questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

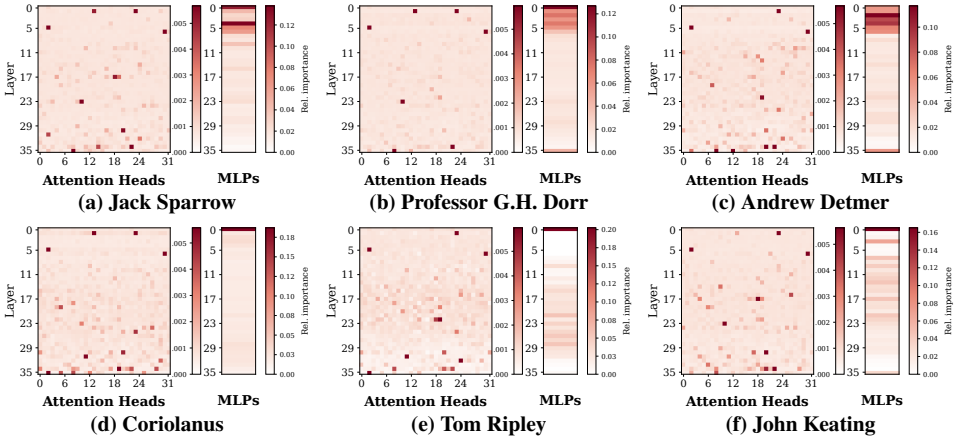



Figure 11: Importance of role-related components in Qwen3-8B-Instruct on role-specific questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

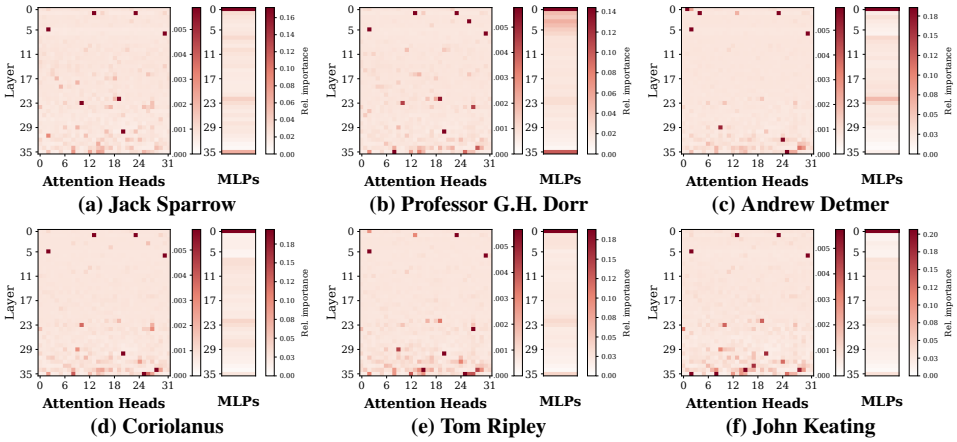



Figure 12: Importance of role-related components in Qwen3-8B-Instruct on general questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

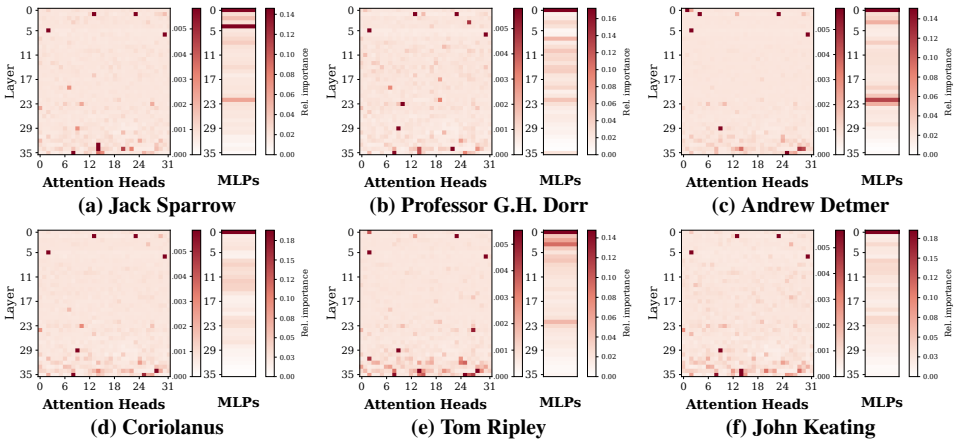



Figure 13: Importance of role-related components in Qwen3-8B-Instruct on general questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

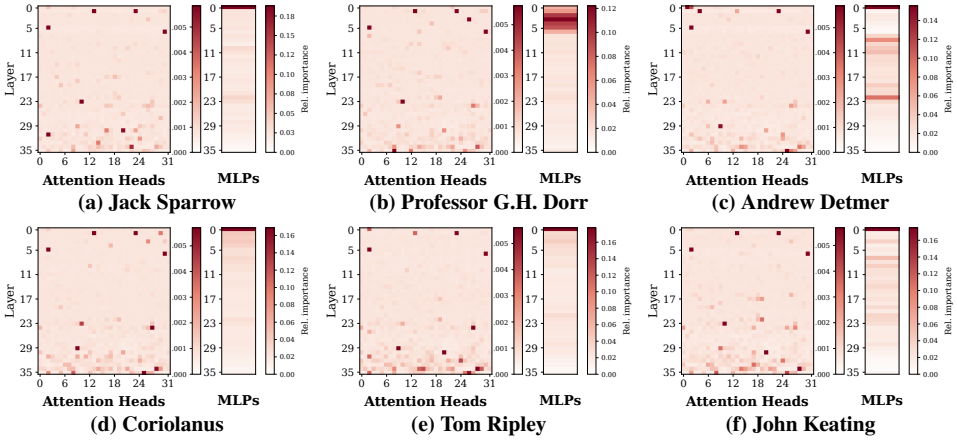



Figure 14: Importance of role-related components in Qwen3-8B-Instruct on general questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

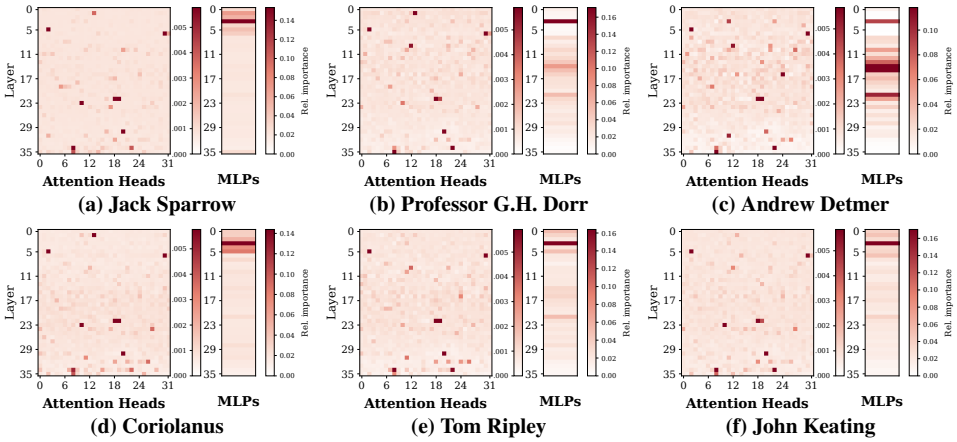



Figure 15: Importance of role-related components in Qwen3-8B Base on role-specific questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

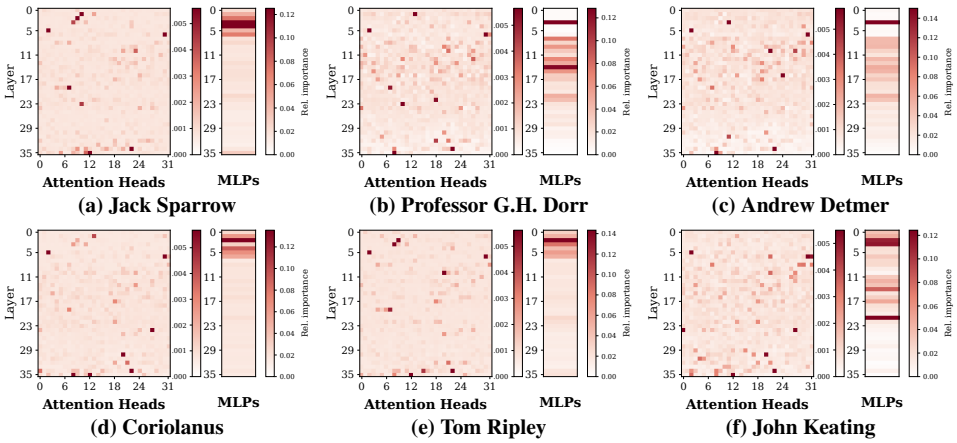



Figure 16: Importance of role-related components in Qwen3-8B Base on role-specific questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

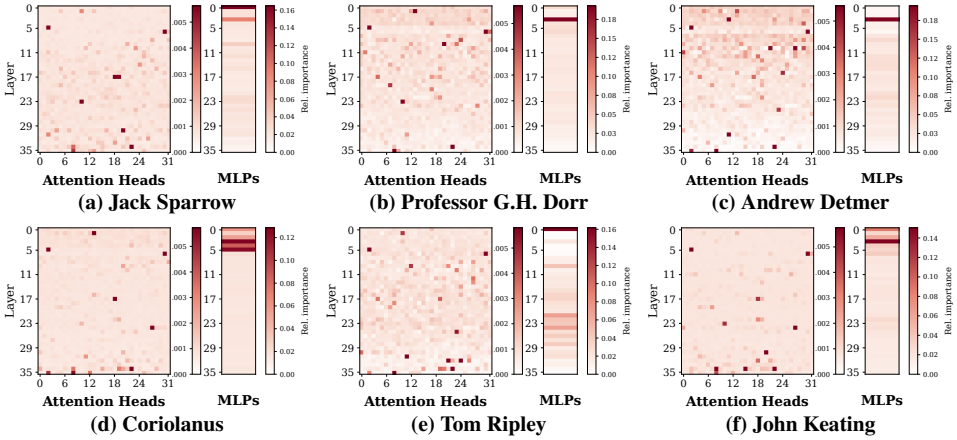



Figure 17: Importance of role-related components in Qwen3-8B Base on role-specific questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

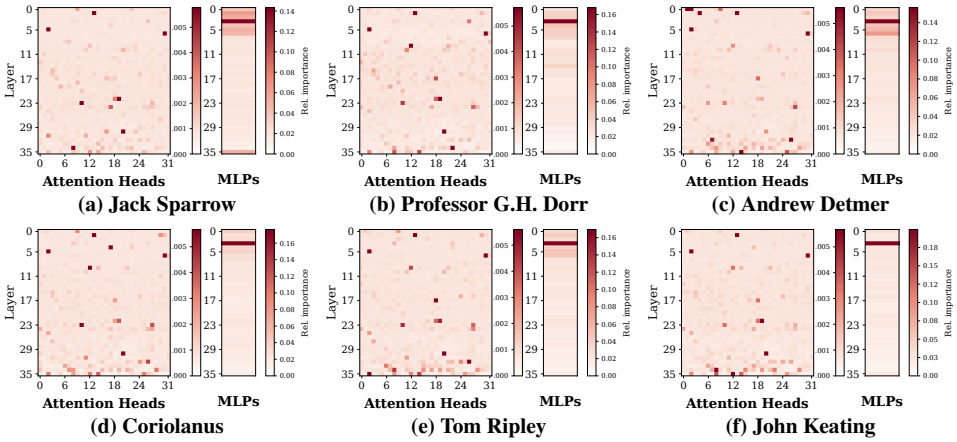



Figure 18: Importance of role-related components in Qwen3-8B Base on general questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

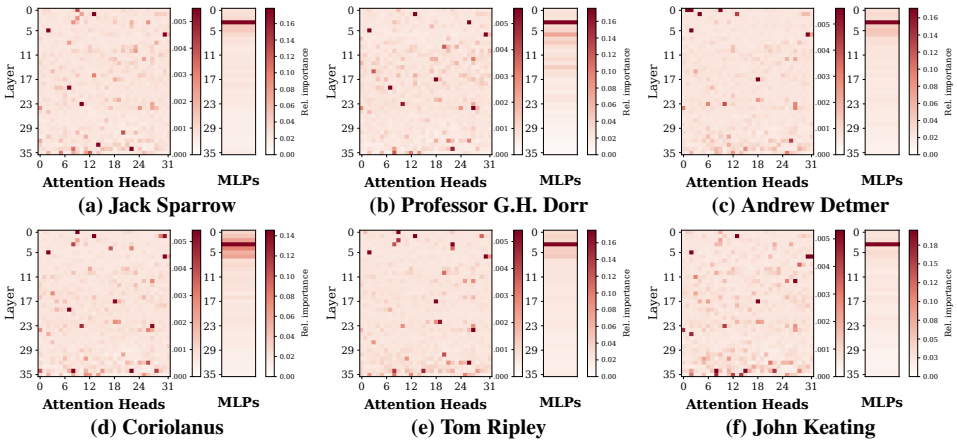



Figure 19: Importance of role-related components in Qwen3-8B Base on general questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

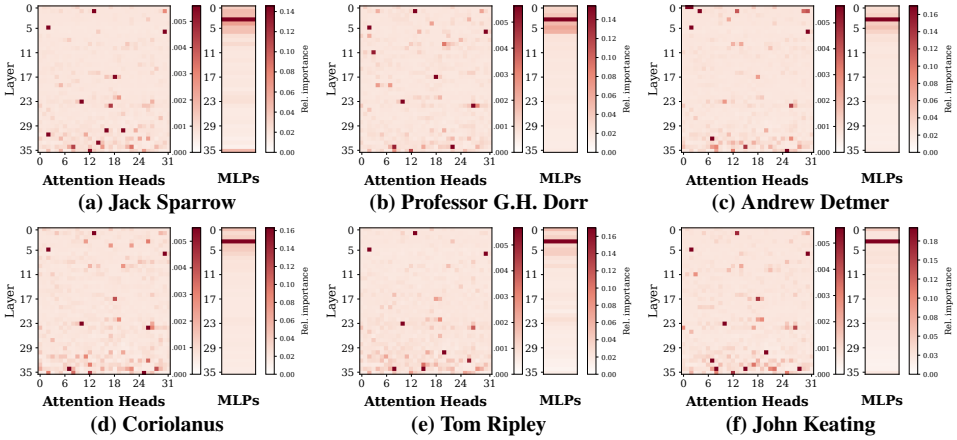



Figure 20: Importance of role-related components in Qwen3-8B Base on general questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

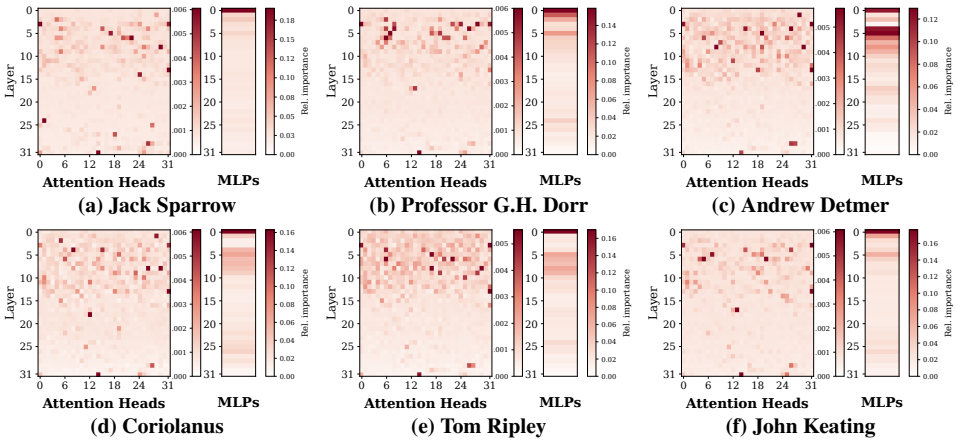



Figure 21: Importance of role-related components in Llama-3.1-8B-Instruct on role-specific questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

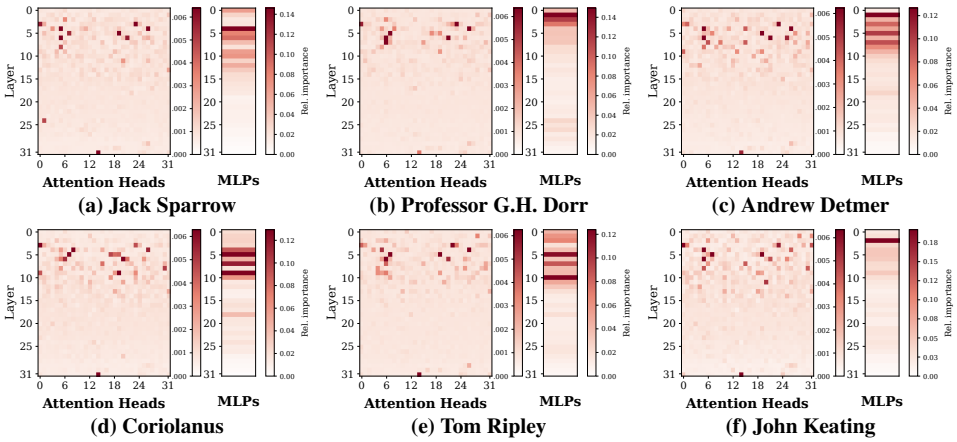



Figure 22: Importance of role-related components in Llama-3.1-8B-Instruct on role-specific questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

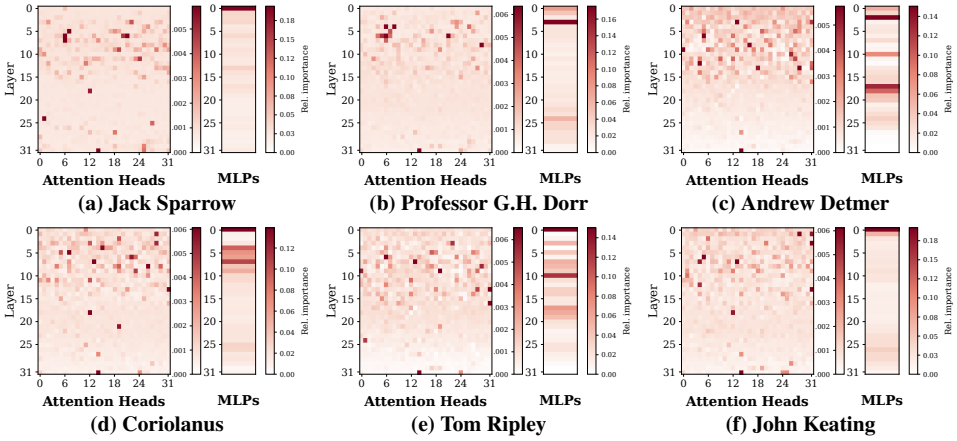



Figure 23: Importance of role-related components in Llama-3.1-8B-Instruct on role-specific questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

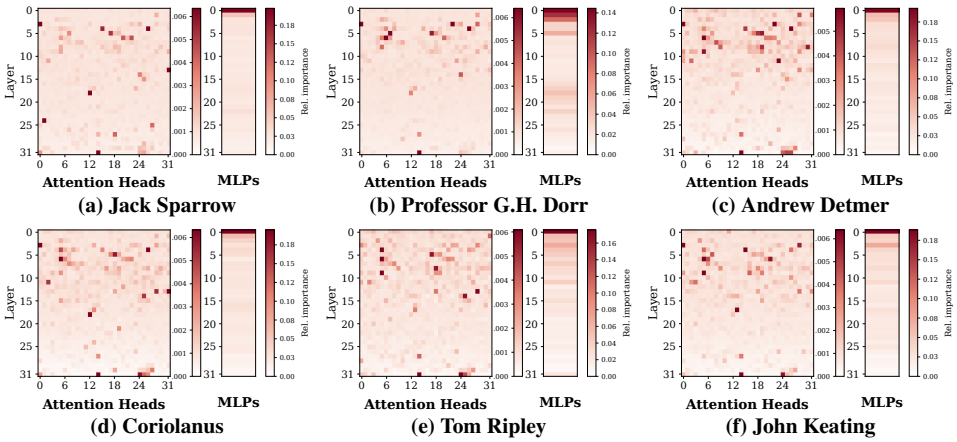



Figure 24: Importance of role-related components in Llama-3.1-8B-Instruct on general questions under the no-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

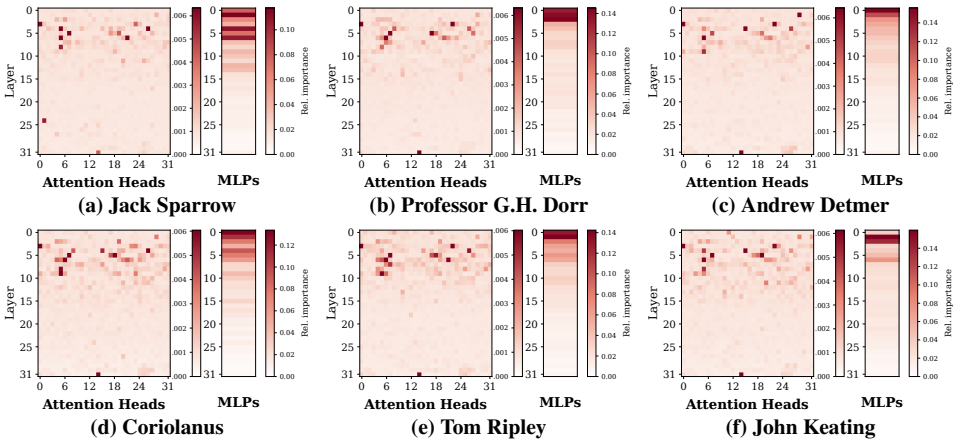



Figure 25: Importance of role-related components in Llama-3.1-8B-Instruct on general questions under the no-profile contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

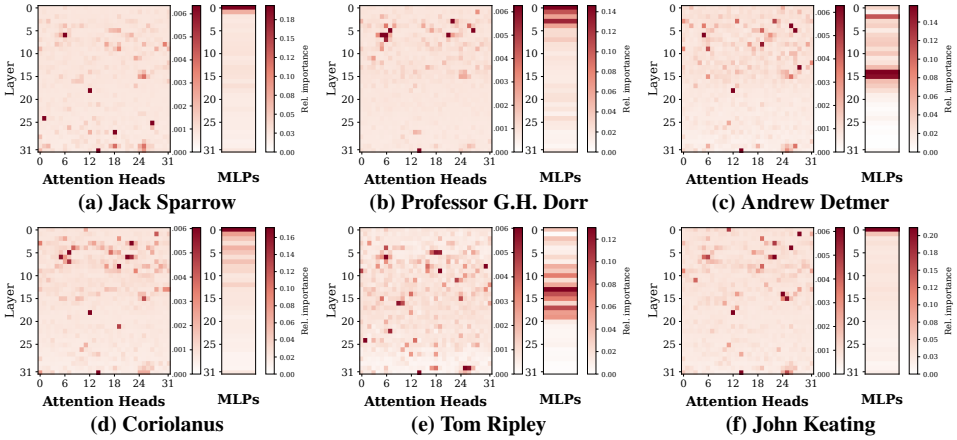



Figure 26: Importance of role-related components in Llama-3.1-8B-Instruct on general questions under the wrong-role contrast. Squares denote attention heads and side strips denote MLP blocks; character panels and component families use independent scales.

### C.4 Extended progressive-knockout results

To test whether the attributed components are causally important for role-playing generation, we compare progressive removal in CoDePass-ranked and random order. We progressively zero attention-head outputs and complete MLP blocks in Qwen3-8B-Instruct. Each attention step removes 20 heads, and each MLP step removes one block. The CoDePass-ranked interventions use the no-role, wrong-role, and no-profile rankings; the control follows a fixed random ordering. All conditions use the same 200 role-specific questions, greedy decoding, a 120-word response constraint, and a maximum of 512 generated tokens. Figure 2 reports ROUGE-L relative to the corresponding unmodified-model baseline.

The unmodified model completes every response before the token limit. Attribution-ranked head interventions increasingly separate from the random ordering as more heads are removed. Ranked MLP interventions separate at smaller intervention sizes, and removing many MLP blocks strongly disrupts generation under every ordering. Table 16 reports the large-intervention endpoints and the accompanying generation failures.

Table 16: Qwen3-8B-Instruct generation after removing 300 attention heads or 15 MLP blocks. ROUGE-L is relative to the unmodified model.

| Ranking | Head ROUGE-L (%) | Head Limit Hits | MLP ROUGE-L (%) | MLP Limit Hits |
|---|---|---|---|---|
| Random | 95.65 | 15 | 23.26 | 200 |
| No role | 18.48 | 188 | 0.00 | 200 |
| Wrong role | 38.41 | 146 | 7.28 | 200 |
| No profile | 38.31 | 166 | 2.43 | 200 |

## C.5 Base-to-Instruct system reorganization

### C.5.1 Representative component ranks

To determine whether the Base-to-Instruct reorganization extends beyond the representative character, we compare component ranks aggregated across characters and prompting conditions. Table 17 extends the Jack Sparrow comparison in Figure 3 to aggregate component ranks. The local score and rank average the six roles in the role-specific/no-role setting. The global columns average all six roles, both scenes, and all three contrasts within each Qwen model.

Table 1 aggregates each component over all characters, scenes, and role contrasts before comparing Base and Instruct. The K–S test compares the resulting component-score distributions: $D = .186$ over 1,152 attention heads and $D = .306$ over 36 MLP blocks. "Changed" counts components replaced from the Base top set, using $k = 32$ for heads and $k = 8$ for MLP blocks. The largest signed shifts are L6H30 at $-6.67$ percentage points and MLP L3 at $-4.29$ points. The unequal family sizes make effect sizes, signed shifts, and top-set overlap more informative than comparing the two $p$-values directly.

Table 17: Representative changes in role-related components from Qwen3-8B Base to Instruct.

| Component | No-Role Score (%) | Local Rank | Global Base $\rightarrow$ Instruct |
|---|---|---|---|
| L5H2 | 4.959 | 1 | $2 \rightarrow 1$ |
| L6H30 | 4.332 | 2 | $1 \rightarrow 3$ |
| L1H23 | 4.286 | 3 | $452 \rightarrow 2$ |
| MLP L0 | 6.811 | 1 | $5 \rightarrow 1$ |
| MLP L35 | 5.911 | 2 | $33 \rightarrow 5$ |

The ranks show both inheritance and rebalancing. L5H2 and L6H30 remain among the three leading global heads, preserving prominent context routes, while L1H23 rises from rank 452 to rank 2; RQ2 identifies its response-interface behavior. The MLP top set changes from the continuous Base sequence L0–L7 to L0–L4 together with L22, L23, and L35. Combining these rank shifts with the RQ2 diagnostics, the Base-to-Instruct system transition redistributes an existing role-sensitive substrate from a front-loaded organization toward coordinated entry, middle-layer identity processing, and final response preparation.

### C.5.2 Emergence of Role-Playing Capability

To trace when role-related components form and how they reorganize during model development, we examine attribution at initialization, after pretraining, and after instruction tuning. We compare three endpoints of the same Qwen3-8B architecture: a fixed random initialization, the pretrained Base checkpoint, and the Instruct checkpoint. The analysis fixes the Jack Sparrow role-specific/no-role condition, the 200 response targets, and the relative-absolute CoDePass estimator. Random and Base use the completion interface, while Instruct uses its chat interface. Top-$k$ mass measures concentration within a component family; the K–S statistic measures distributional change between endpoints.

Table 18 shows the clearest stage transition in attention. From random initialization to Base, the mass carried by the top 32 heads triples and the largest component score rises from .55% to 19.49%. The Instruct endpoint remains substantially more concentrated than a uniform head distribution

Table 18: Concentration of role-related components across random, Base, and Instruct systems.

| System | Components | Top-$k$ Mass (%) | Uniform Mass (%) | Maximum Score (%) |
|---|---|---|---|---|
| Random | Attention heads ($k = 32$) | 10.29 | 2.78 | .55 |
| Base | Attention heads ($k = 32$) | 30.89 | 2.78 | 19.49 |
| Instruct | Attention heads ($k = 32$) | 15.69 | 2.78 | 3.53 |
| Random | MLP blocks ($k = 8$) | 41.08 | 22.22 | 6.75 |
| Base | MLP blocks ($k = 8$) | 44.72 | 22.22 | 11.17 |
| Instruct | MLP blocks ($k = 8$) | 29.38 | 22.22 | 5.91 |

while spreading mass more broadly than Base, marking a second stage of reorganization. The MLP concentration changes much less from random initialization to Base.

Table 19: Changes in role-related components across training stages.

| Transition | Components | Top-$k$ Overlap | K–S Statistic | $p$-Value |
|---|---|---|---|---|
| Random → Base | Attention heads | 0 of 32 | .252 | $1.8\times10^{-32}$ |
| Random → Base | MLP blocks | 5 of 8 | .194 | .510 |
| Base → Instruct | Attention heads | 17 of 32 | .584 | $2.0\times10^{-182}$ |
| Base → Instruct | MLP blocks | 6 of 8 | .556 | $1.9\times10^{-5}$ |

None of the random model's leading attention heads remains in the Base top set, whereas the Instruct system retains more than half of the Base heads and six of its eight leading MLP blocks (Table 19). The endpoint sequence reveals two stages of role-related attention organization: a concentrated set forms during pretraining, and instruction tuning inherits and redistributes this set. The MLP transition occurs later, from Base to Instruct, as the attention scaffold becomes coupled to the transformation pathway.

### C.6 Attribution stability with fewer examples

To assess the sensitivity of component discovery to attribution sample size, we examine how many examples are needed to recover the shared attention-head ranking. For eight training characters in both scenes, we compute Qwen3-8B-Instruct reference rankings using 200 examples per character–scene cell. We then sample 16, 32, 48, or 64 examples without replacement within each cell and average the 16 cell matrices with equal weight, repeating this procedure 1,000 times. Table 20 reports the mean Spearman correlation with the reference ranking and Top-64 overlap, defined as the number of shared heads divided by 64.

Table 20: Stability of shared attention-head rankings with fewer attribution examples. Values average 1,000 subsamples relative to a 200-example-per-cell reference.

| Examples per Cell | Spearman Correlation | Top-64 Overlap (%) |
|---|---|---|
| 16 | 0.9766 | 89.95 |
| 32 | 0.9883 | 92.64 |
| 48 | 0.9926 | 94.04 |
| 64 | 0.9949 | 95.08 |

With only 32 examples per cell, the shared ranking reaches a correlation of 0.9883 and retains 92.64% of the reference Top-64 heads. Increasing the sample count to 64 raises overlap to 95.08%. The high agreement at 32 examples shows that the shared high-ranking components can be recovered from a small attribution set, supporting the sample count used for targeted fine-tuning.

# D BEHAVIORAL DIAGNOSTICS AND TRANSFER ANALYSIS

This appendix specifies the diagnostic positions, token regions, probes, and frozen class criteria used in RQ2, then tests whether the resulting head functions and MLP trajectories transfer across roles and model architectures.

## D.1 ATTENTION-HEAD DIAGNOSTIC PROTOCOL

### D.1.1 RESPONSE-PREPARATION POSITION

The Qwen chat template is rendered with thinking disabled. The assistant scaffold contains an empty `<think>...</think>` block followed by the final-answer boundary. The last scaffold token is used as the response-preparation query in Equation 3. Attention maps and MLP states are computed from the prompt before any answer token is appended.

### D.1.2 ATTENTION TOKEN REGIONS

Four regions partition the rendered prompt: (i) the role instruction before the profile marker, (ii) profile content, (iii) the user question, and (iv) chat-template scaffold tokens. Additional regions mark every intended-role occurrence, the role-indicator phrases, and the final eight scaffold tokens. We analyze both the complete $N \times N$ token-interaction matrix and its response-preparation row, capturing vertical token-selection patterns and diagonal positional patterns. Qwen uses grouped-query attention. For each query head, the value-vector norm is taken from its corresponding key/value head before output projection. Equation 3 therefore measures $A_{qk}\|V_k\|_2$ for each token interaction.

### D.1.3 SAMPLING AND FROZEN-CLASS CRITERIA

The attention analysis selects the top 20 role-specific/no-role heads after averaging attribution scores across characters and applies the same candidates to both scenes. Complete interaction maps yield three explicit criteria: identity-token share $\geq .20$ for role-indicator heads, self-plus-previous-token mass $\geq .50$ for positional heads, and profile-plus-question share $\geq .55$ for role-context heads. The criteria are defined once on 100 Jack Sparrow role-specific prompts and classify seven Qwen heads: three context, two indicator, and two positional. Class membership is transferred unchanged to five held-out characters in both role-specific and general-question scenes, with 100 prompts in each character–scene cell. Distribution plots average heads within each class and report the mean and sample standard deviation over the 100 examples.

## D.2 TRANSFER OF ATTENTION-HEAD FUNCTIONS

### D.2.1 SCENE REALLOCATION WITHIN THE DISCOVERY ROLE

To test whether head functions persist when answers depend less on character knowledge, we compare the frozen classes between role-specific and general questions. Figure 5 shows that the functional operation persists while information use follows the scene. Context heads exchange profile mass for question mass on general prompts, indicator heads reduce identity concentration, and positional heads retain their response-boundary pattern.

To assess whether these functions extend beyond the discovery character, we apply the same head classes to five held-out characters in both scenes. Figure 27 preserves the qualitative separation of the three classes. Across the ten held-out character–scene conditions, role-context and positional heads satisfy their original criteria in 10/10 conditions, with mean scores of .723 and .646. Role-indicator heads satisfy the character-name criterion in 5/10 conditions, with a mean score of .223. Context routing and position therefore transfer stably, while identity selection exhibits partial transfer.

### D.2.2 CROSS-ARCHITECTURE TRANSFER

To test whether the same functional organization appears in another architecture, we transfer the classification criteria to Llama and evaluate its selected heads across characters. We repeat the frozen-class analysis with Llama-3.1-8B-Instruct using its own role-specific/no-role ranking. The same criteria applied to 100 Jack Sparrow role-specific prompts identify two context, two indicator,

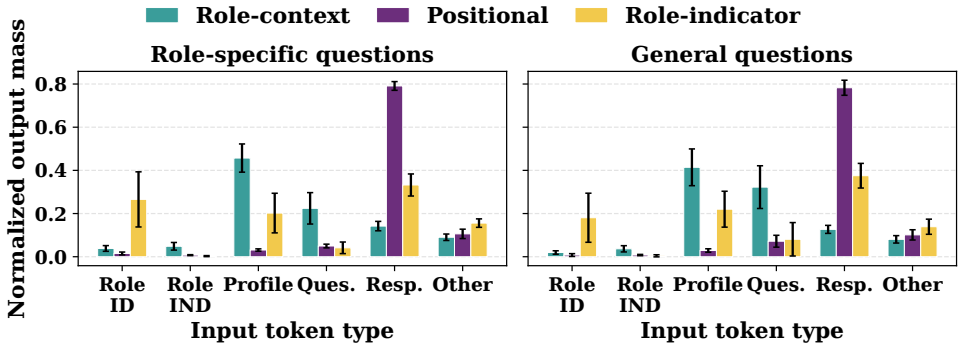



Figure 27: Cross-character transfer of the three attention-head classes. ID and IND denote identity and role-indicator tokens; Resp. denotes the response boundary.

and one positional head. Formal transfer covers Coriolanus, Tom Ripley, and John Keating in both scenes, with 100 prompts per character–scene cell.

Table 21: Cross-architecture transfer of the three attention-head classes.

| Model | Context | Indicator | Positional |
|---|---|---|---|
| Qwen3-8B-Instruct | 10/10 | 5/10 | 10/10 |
| Llama-3.1-8B-Instruct | 6/6 | 1/6 | 6/6 |

Table 21 shows that context and positional functions transfer in every formal condition for both architectures, whereas identity selection transfers only partially. The shared prompt schema supports stable content and timing operations; binding a specific character remains persona- and architecture-dependent.

### D.3 MLP PROBES AND CONTROLS

To distinguish information entering an MLP from the information it adds, we probe the incoming residual state and MLP update separately for identity, profile, role instruction, and response content. The intended identity, positive profile, role-indicator phrase, and reference response define four signal probes. A fifth probe uses the fixed wrong-role identity as a semantic control. Each signal probe is truncated to at most 64 tokens and paired with an equal-length sample from a fixed pool of ordinary neutral English tokens. Probe vectors are constructed from rows of the model's output embedding matrix, and each role-level curve averages 100 prompts.

To test whether semantic answer formation precedes persona-specific expression, we use general-instruction questions with meaning-matched answers in neutral and persona-rendered forms. A neutral answer to a different instruction serves as the unrelated control. The Qwen analysis uses 100 general-instruction prompts for each of the six characters. Each response direction is the mean of at most 64 output-embedding rows. All three directions are evaluated on the positive role prompt at the response-preparation position. We summarize the middle and late stages over L18–L28 and L29–L35.

### D.4 MLP TRAJECTORIES ACROSS ROLES AND ARCHITECTURES

**Qwen MLP trajectories.** To determine whether semantic response formation and persona modulation follow a shared sequence across characters, we compare layerwise alignment with meaning-matched neutral and persona-rendered answers and an unrelated control.

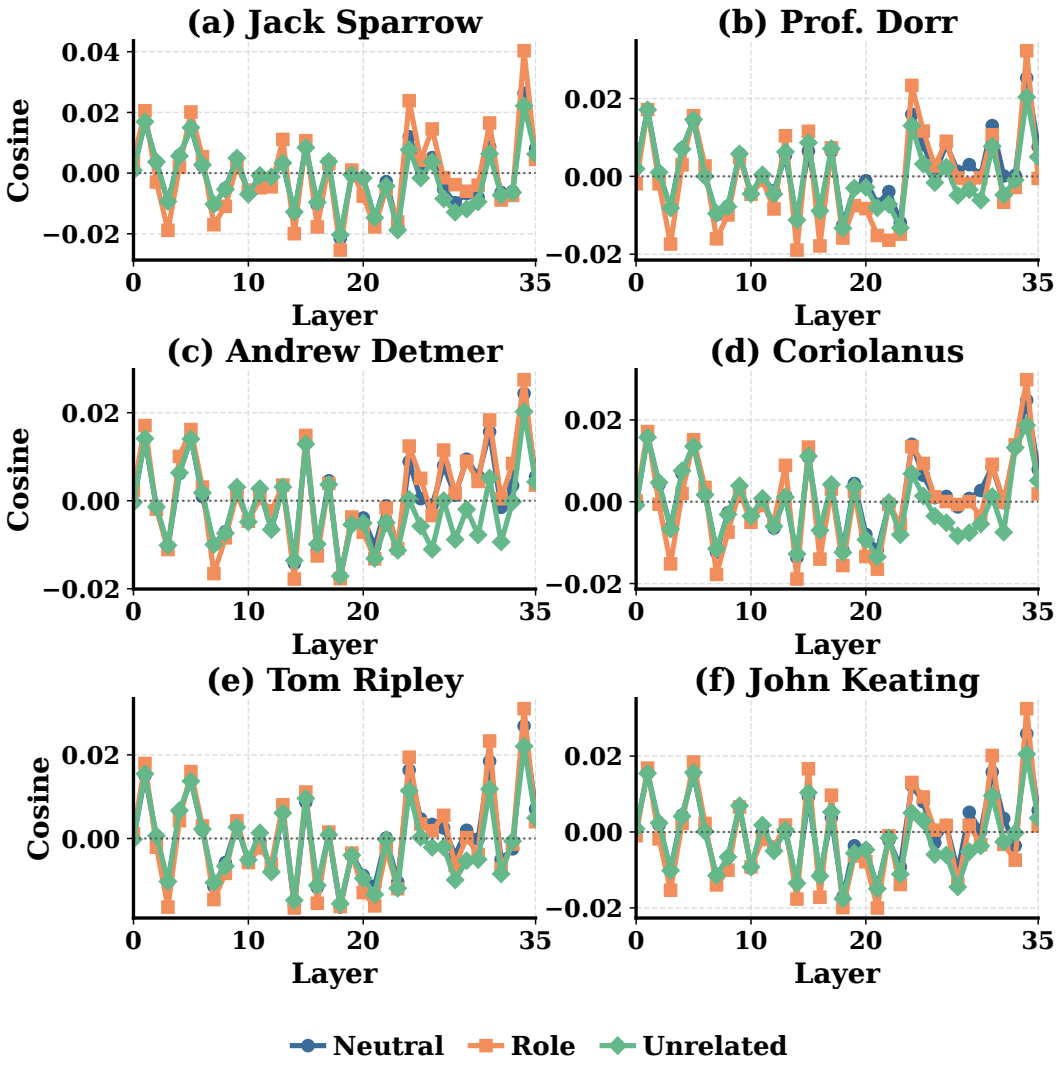



Figure 28: MLP-update alignment with response directions across six characters.

Figure 28 shows a shared layerwise geometry for neutral, persona-rendered, and unrelated response directions. Within the middle band, both meaning-matched directions exceed the unrelated direction on average for every role, while the neutral–persona difference remains much smaller. Jack Sparrow and Andrew Detmer keep the persona direction at or above the neutral direction, with neutral-minus-persona differences of $-.0019$ and $-.0003$. Professor Dorr, Coriolanus, Tom Ripley, and John Keating show small neutral leads of $+.0024$, $+.0012$, $+.0011$, and $+.0014$. Both trajectory families share semantic response formation and differ in the timing of persona modulation.

**Llama MLP trajectories.** To test whether the Qwen transformation sequence recurs in another architecture, we apply the same response-direction probes to Llama. The Llama analysis uses 100 general-instruction prompts for each of Jack Sparrow, Coriolanus, Tom Ripley, and John Keating.

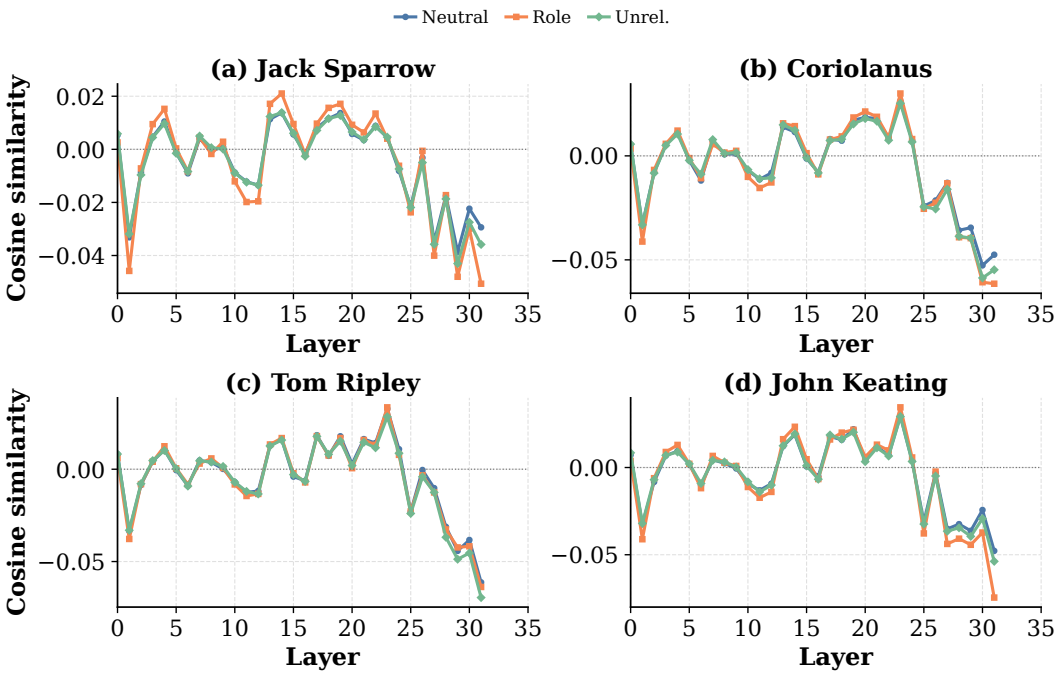



Figure 29: Llama-3.1-8B-Instruct MLP-update alignment with response directions across four characters.

The three Llama directions follow nearly identical layerwise geometry in Figure 29, and none of the four characters exhibits the Qwen pattern of a middle-layer neutral lead followed by a late persona lead. Llama thus preserves the functional split in content and positional routing but realizes late response transformation differently. Across the two architectures, the stable result is reusable prompt routing; identity selection and persona-sensitive MLP readout are the variable part of the mechanism.

# E Targeted Fine-Tuning: Experimental Details and Additional Results

## E.1 Experimental setup and efficiency

**Component selection.** The main experiment aligns component discovery with supervised training and separates both from evaluation. We take the 75 characters shared by the role-specific and general training splits, estimate a CoDePass matrix for each character in each scene, and average all 150 matrices with equal weight. Each matrix uses 32 matched examples, a sample count chosen for its high agreement with 200-example rankings (Appendix C.6). The six evaluation characters are excluded from this 75-character set. We select the 64 highest-ranked query-head groups and update their required Q/K/V/O projection slices. Under grouped-query attention, multiple query heads can share a K/V group, which is updated once. This selection trains 124,780,544 parameters. Random controls match its layerwise query-head counts, K/V groups, and total parameter count.

**Training setup.** The SFT corpus contains 18,205 training and 994 development examples from the same 75 characters, balanced across role-specific and general scenes with at most 128 examples per character–scene cell. All Qwen variants train for two epochs with global batch size 64, AdamW (Loshchilov & Hutter, 2019), cosine decay, 2% warmup, maximum sequence length 1,024, and no thinking output. The three main methods are trained with seeds 13, 21, and 42. All SFT methods uses the learning rate $10^{-4}$.

**Evaluation.** For each of the six evaluation characters and each scene, we use the first 100 RoleBench test questions. Generation is greedy, thinking is disabled, and responses are constrained to at most

120 words. The primary metric is ROUGE-L against RoleBench's first model-generated reference, matching the benchmark protocol; SacreBLEU is secondary. MMLU uses 5-shot accuracy, ARC-Challenge 25-shot normalized accuracy, and HellaSwag 10-shot normalized accuracy. Instruction models use their chat templates; Base models use their plain-text prompt interfaces.

**Result reporting.** Evaluation scores use a 0–100 scale; ± denotes the sample standard deviation across training seeds where reported. Training time is the mean wall time on eight H100 GPUs divided by the training-set size times the number of epochs, in milliseconds per example, excluding attribution and evaluation. The main experiment uses 18,205 examples. Tuned parameter counts use B for billions and M for millions, with the percentage of the full model in parentheses. Bold entries mark the best value within each comparison group, including ties at the displayed precision.

**Training efficiency.** Targeted SFT takes 32.3 ms per training example, compared with 50.3 ms for Full SFT, a 35.9% reduction. Random SFT updates the same number of parameters and takes 32.9 ms per example. The smaller trainable set reduces optimization cost while preserving the full forward computation.

### E.2 Results on additional models

To examine how targeted fine-tuning generalizes across architectures and pretrained or instruction-tuned starting points, we evaluate three additional models in Table 22. We apply the same 75-character, dual-scene discovery and training split to Llama-3.1-8B-Instruct, Llama-3.1-8B-Base, and Qwen3-8B-Base, and evaluate the same six unseen characters. Llama uses a 5.00% trainable-parameter budget for Targeted and matched Random SFT, with one training epoch; Qwen Base uses the Top-64 selection (1.49%) and two epochs.

Table 22: Evaluation on role-playing and generic tasks with additional models. ± denotes the sample standard deviation over three training seeds.

| Models | Train Time (ms/sample)↓ | Tuned Parameters↓ | Role-Playing Tasks | | | | Generic Tasks | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Role-Specific | | General Scene | | MMLU | ARC | HellaSwag |
| | | | ROUGE-L↑ | SacreBLEU↑ | ROUGE-L↑ | SacreBLEU↑ | Acc.↑ | Acc.↑ | Acc.↑ |
| Llama-3.1-8B-Base | – | – | 13.30 | 2.28 | 12.55 | 1.16 | **65.35** | 58.45 | 81.95 |
| + Full SFT | 43.0 | 8.03B (100%) | 22.67±0.93 | 4.40±1.03 | 36.28±0.73 | 9.47±0.31 | 52.74±1.14 | 54.89±0.88 | 79.89±0.14 |
| + Random SFT | 32.3 | **401.6M (5.00%)** | 24.92±0.10 | 6.77±0.09 | 35.31±0.41 | 12.17±0.39 | 62.32±1.12 | 61.40±0.74 | 82.00±0.22 |
| **+ Targeted SFT** | **31.2** | **401.6M (5.00%)** | **25.33±0.06** | **7.46±0.18** | **36.89±0.56** | **12.47±0.33** | 64.82±0.34 | **61.49±0.25** | **82.40±0.26** |
| Llama-3.1-8B-Instruct | – | – | 18.25 | 3.39 | 11.50 | 2.68 | **68.67** | **64.25** | **78.35** |
| + Full SFT | 42.5 | 8.03B (100%) | **24.24±0.27** | 5.28±0.29 | 35.71±0.49 | 11.96±0.46 | 65.75±1.45 | 58.30±0.10 | 76.98±0.36 |
| + Random SFT | 31.4 | **401.6M (5.00%)** | 24.04±0.04 | 5.81±0.04 | 32.68±0.29 | 11.02±0.16 | 66.48±0.03 | 61.82±0.09 | 75.64±0.02 |
| **+ Targeted SFT** | **30.9** | **401.6M (5.00%)** | **24.24±0.20** | **6.05±0.29** | **36.10±0.42** | **12.84±0.18** | 67.39±0.40 | 62.17±0.52 | 77.34±0.49 |
| Qwen3-8B-Base | – | – | 20.85 | 5.46 | 20.45 | 3.65 | **76.78** | 67.49 | 79.50 |
| + Full SFT | 47.2 | 8.19B (100%) | 23.81±0.12 | **7.17±0.15** | 34.13±0.05 | 10.69±0.60 | 76.53±0.02 | 67.55±0.21 | 79.40±0.13 |
| + Random SFT | **32.3** | **121.6M (1.49%)** | 23.84±0.32 | 6.47±0.25 | 36.16±0.23 | 12.32±0.51 | 76.56±0.13 | 67.97±0.32 | 79.47±0.10 |
| **+ Targeted SFT** | 32.4 | **121.6M (1.49%)** | **25.14±0.27** | 7.12±0.27 | **38.33±0.64** | **14.16±0.30** | 76.76±0.11 | **68.43±0.39** | **80.31±0.16** |

**Role-playing performance.** On Llama Base, Targeted SFT attains the highest role-specific ROUGE-L and SacreBLEU, improving ROUGE-L by 1.7% over Random SFT and 11.7% over Full SFT. On Llama Instruct, Targeted SFT improves role-specific and general-scene ROUGE-L over Random SFT by 0.8% and 10.5%. On Qwen Base, Targeted SFT improves ROUGE-L over Random and Full SFT in both scenes. And, Targeted SFT gains 5.4% and 5.6% on role-specific questions and 6.0% and 12.3% on general questions, respectively.

**General capabilities and efficiency.** Targeted SFT preserves substantially more of Llama Base's MMLU performance than Full SFT (64.82 versus 52.74) and reaches the highest HellaSwag score. For Llama Instruct, Targeted SFT retains the highest generic-task scores among the adapted models. Targeted SFT trains only 5.00% of each Llama model and reduces training time by 27.3% and 27.4% relative to Full SFT for Instruct and Base, respectively. On Qwen Base, Targeted SFT leads ARC and HellaSwag while retaining MMLU near the original model's level; it updates 1.49% of parameters and reduces training time by 31.3% relative to Full SFT.

### E.3 Consistency across random component selections

To test whether the advantage of Targeted SFT depends on the particular random components selected, we compare three independently drawn, capacity-matched Random Top-64 sets and Full SFT. All models answer the same questions for the six held-out characters. Table 23 reports the paired ROUGE-L differences in both scenes.

Table 23: ROUGE-L gains of Targeted SFT on held-out characters.

| Comparison | Role-Specific Δ | General Δ |
|---|---|---|
| Targeted − Random Set 1 | +2.08±0.20 | +5.01±0.85 |
| Targeted − Random Set 2 | +1.61±0.10 | +6.39±0.55 |
| Targeted − Random Set 3 | +1.63±0.09 | +8.71±0.14 |
| Targeted − Full SFT | +0.97±0.20 | +3.16±0.18 |

**Role-specific gains.** Targeted SFT exceeds the three Random sets by 2.08, 1.61, and 1.63 ROUGE-L points and Full SFT by 0.97 points. The ordering persists across the independently selected random coordinates, supporting the value of the CoDePass ranking beyond parameter count alone.

**Variation in random component selection.** The same selected components improve general-scene ROUGE-L over all three Random sets by 5.01–8.71 points and Full SFT by 3.16 points. Thus, the shared ranking supports both profile-dependent and general instructions on unseen characters.

### E.4 Capacity sweep and MLP ablation

To determine how many attention heads need to be fine-tuned to improve role playing and whether MLP updates provide additional gains, we vary the number of selected heads and augment them with complete MLP blocks.

**Attention capacity and response length.** Increasing the attribution-selected QKVO capacity from 8 to 64 heads improves role-specific ROUGE-L from 21.25 to 23.39 (Table 24). To determine whether this gain comes from more precise content selection or longer answers, we compare precision, recall, and response length (Table 6). QKVO-64 raises precision to 39.49 and attains recall of 17.22, close to the five-head value of 17.63, while reducing mean response length from 52.9 to 28.9 words. Broader ranked coverage therefore improves reference-content selection without relying on longer answers.

Table 24: Extended attention and MLP capacity controls on held-out characters.

| Trainable Set | Train Time (ms/sample)↓ | Tuned Parameters↓ | Role-Playing Tasks | | | | Generic Tasks | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Role-Specific | | General Scene | | MMLU | ARC | HellaSwag |
| | | | ROUGE-L↑ | SacreBLEU↑ | ROUGE-L↑ | SacreBLEU↑ | Acc.↑ | Acc.↑ | Acc.↑ |
| QKVO-8 | **28.4** | **16.8M (0.20%)** | 21.25±0.28 | 3.43±0.31 | 34.18±0.50 | 11.01±0.94 | 72.51±0.36 | 65.02±0.23 | 73.40±0.17 |
| QKVO-16 | 29.4 | 32.5M (0.40%) | 21.87±0.20 | 3.94±0.35 | 35.89±0.78 | 11.94±0.45 | 72.57±0.71 | 66.27±1.31 | 74.95±0.21 |
| QKVO-32 | 29.9 | 65.0M (0.79%) | 22.95±0.21 | 3.86±0.18 | 37.64±0.57 | 12.80±0.24 | 73.57±0.09 | 65.81±0.38 | 74.08±0.33 |
| QKVO-64 | 32.3 | 124.8M (1.52%) | 23.39±0.11 | 4.32±0.06 | 39.08±0.12 | **13.34±0.19** | **74.21±0.12** | 65.93±0.32 | 75.16±0.31 |
| QKVO-96 | 34.2 | 186.6M (2.28%) | 23.27±0.17 | 4.38±0.14 | 39.00±0.08 | 12.90±0.59 | 73.85±0.44 | 65.42±0.91 | 75.15±0.14 |
| QKVO-128 | 33.8 | 242.2M (2.96%) | 23.46±0.05 | 4.40±0.16 | 39.32±0.35 | 12.77±0.40 | 73.55±0.39 | 64.93±0.09 | 75.31±0.10 |
| QKVO-160 | 34.2 | 296.7M (3.62%) | 23.44±0.13 | 4.52±0.07 | **39.46±0.29** | 12.78±0.20 | 73.42±0.50 | 64.59±0.39 | 75.23±0.17 |
| QKVO-64 + Top-1 MLP | 32.7 | 275.8M (3.37%) | **23.70±0.23** | **4.76±0.16** | 38.74±0.62 | 12.99±0.43 | 73.89±0.33 | 65.42±0.26 | 75.14±0.29 |
| QKVO-64 + Top-2 MLPs | 32.8 | 426.8M (5.21%) | 23.13±0.17 | 4.23±0.20 | 38.79±0.35 | 12.25±0.25 | 73.57±0.45 | 64.65±0.77 | 74.74±0.50 |
| QKVO-64 + Top-3 MLPs | 32.3 | 577.8M (7.05%) | **23.70±0.30** | 4.74±0.25 | 39.12±0.41 | 12.45±0.49 | 73.39±0.43 | 63.25±0.49 | 74.85±0.09 |
| QKVO-64 + Top-5 MLPs | 33.5 | 879.8M (10.74%) | 22.70±0.40 | 3.76±0.30 | 38.85±0.39 | 11.75±0.69 | 73.00±0.35 | 62.14±0.74 | 75.35±0.32 |
| QKVO-64 + All MLPs | 50.2 | 5.56B (67.89%) | 21.47±0.38 | 3.68±0.39 | 37.70±0.42 | 11.24±0.46 | 62.65±0.99 | 48.27±2.64 | 67.39±0.59 |
| QKVO-64 + Random-3 MLPs | 33.2 | 577.8M (7.05%) | 23.62±0.36 | 4.73±0.29 | 38.90±0.26 | 12.35±0.09 | 74.20±0.27 | 64.14±0.42 | **75.85±0.29** |

**Extended attention capacity.** To test whether the gains saturate beyond the Top-64 selection, we extend the attention sweep and compare broader MLP updates. Extending the ranked attention set from 64 to 96–160 query-head groups leaves role-specific ROUGE-L between 23.27 and 23.46, compared with 23.39 for Top-64, while general-scene ROUGE-L ranges from 39.00 to 39.46 (Table 24). The highest-ranked groups capture most of the role-specific benefit; adding lower-ranked groups produces only small changes on general questions.

The remaining metrics show a trade-off rather than a uniform gain from increased capacity. Across the 96–160-head settings, general-scene SacreBLEU stays below Top-64 and MMLU ranges from 73.42 to 73.85, compared with 74.21 for Top-64. Increasing the number of trainable heads therefore does not consistently improve the joint role-playing and generic-task results.

**MLP augmentation.** Selective MLP updates provide a smaller, task-dependent benefit. Adding the top block increases role-specific ROUGE-L from 23.39 to 23.70 and SacreBLEU from 4.32 to 4.76, but lowers general-scene ROUGE-L from 39.08 to 38.74. The top three blocks yield role-specific ROUGE-L of 23.70, compared with 23.62 for three random blocks (Table 24). Top-2 and Top-5 augmentation yield 23.13 and 22.70, both below attention-only tuning on role-specific ROUGE-L.

Updating all 36 MLP blocks lowers role-specific ROUGE-L to 21.47 and general-scene ROUGE-L to 37.70, with MMLU, ARC, and HellaSwag falling to 62.65, 48.27, and 67.39 (Table 24). Combined with RQ2's staged trajectories, these results distinguish the MLP pathway's role in response construction from the value of updating its weights. Attention-only tuning can improve role playing through an already organized transformation pathway; selective MLP refinement can improve particular response metrics, whereas broad MLP updates disrupt the balance across tasks.

### E.5 Functional-class controls

To identify which routing functions account for the tuning gains, we compare individual functional classes and their combinations while varying the updated projections. The functional-class experiments compare role-context, role-indicator, and positional selections with a matched random set under the shared training and held-out-character evaluation protocol (Table 25).

Table 25: Performance and efficiency of functional-head and projection-slice controls.

| Trainable Set | Train Time (ms/sample)↓ | Tuned Parameters↓ | Role-Playing Tasks | | | | Generic Tasks | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Role-Specific | | General Scene | | MMLU | ARC | HellaSwag |
| | | | ROUGE-L↑ | SacreBLEU↑ | ROUGE-L↑ | SacreBLEU↑ | Acc.↑ | Acc.↑ | Acc.↑ |
| Qwen3-8B-Instruct | – | – | 18.09 | 3.30 | 11.84 | 2.68 | 71.94 | 63.31 | 55.87 |
| Context Q/O | 27.3 | 3.1M (0.04%) | 19.93±0.04 | 4.58±0.06 | 24.88±0.38 | 6.00±0.29 | 71.69±0.26 | 63.91±1.04 | 67.49±0.53 |
| Indicator Q/O | **26.5** | **2.1M (0.03%)** | 22.33±0.05 | 4.04±0.06 | 31.65±0.04 | 9.71±0.34 | 73.02±0.04 | 64.87±0.30 | 65.57±0.26 |
| Positional Q/O | 26.6 | **2.1M (0.03%)** | 21.80±0.08 | 3.56±0.02 | 32.03±0.01 | 8.96±0.11 | 73.00±0.04 | 65.93±0.18 | 68.55±0.38 |
| Context + Indicator Q/O | 26.9 | 5.2M (0.06%) | 21.05±0.28 | **5.37±0.04** | 29.59±1.10 | 7.05±0.52 | 72.49±0.13 | 64.76±0.37 | 69.72±0.80 |
| All Functional Q/O | 27.0 | 7.3M (0.09%) | 21.51±0.36 | 4.64±0.27 | 32.31±0.04 | 8.95±0.11 | 72.59±0.11 | 65.59±0.38 | 70.66±0.24 |
| Context + Indicator QKVO | 27.2 | 10.5M (0.13%) | 20.51±0.32 | 3.62±0.08 | 33.93±0.12 | 11.08±1.27 | 72.59±0.21 | **65.98±0.38** | 72.91±0.23 |
| Matched Random Q/O | 26.9 | 5.2M (0.06%) | 20.89±0.03 | 3.74±0.05 | 29.86±0.25 | 8.87±0.14 | 71.54±0.08 | 63.11±0.61 | 68.17±0.39 |
| Full SFT | 50.3 | 8.19B (100%) | 22.42±0.11 | 4.28±0.09 | 35.91±0.20 | 11.36±0.71 | 73.85±0.21 | 65.90±0.18 | 73.85±0.09 |
| QKVO-64 | 32.3 | 124.8M (1.52%) | **23.39±0.11** | 4.32±0.06 | **39.08±0.12** | **13.34±0.19** | **74.21±0.12** | 65.93±0.32 | **75.16±0.31** |

Table 25 shows that indicator heads attain the highest role-specific ROUGE-L among the functional Q/O selections (22.33), while context+indicator heads attain the highest SacreBLEU (5.37). Combining all three classes does not improve on the indicator selection. Identity selection is therefore an effective role-control point, but the benefits of different routing functions do not simply add together.

Extending context+indicator tuning from Q/O to QKVO lowers role-specific ROUGE-L from 21.05 to 20.51 while raising general-scene ROUGE-L from 29.59 to 33.93 and SacreBLEU from 7.05 to 11.08. K/V updates improve general instruction following within this set, whereas broader Top-64 coverage is more effective for role-specific responses.

The generic-task results distinguish the same interventions: context+indicator QKVO attains the highest ARC accuracy, while Top-64 tuning leads MMLU and HellaSwag. The functional selections take 26.5–27.3 ms per training example, compared with 50.3 ms for Full SFT and 32.3 ms for Top-64 tuning, trading a smaller update budget for lower role-specific ROUGE-L.

### E.6 Qualitative case analysis

To understand how targeted fine-tuning improves role-playing responses and where further progress is needed, we compare automatic response indicators and matched successful and unsuccessful examples across the original and fine-tuning models. We compare Qwen3-8B-Instruct, Targeted SFT, three matched Random SFT models, and Full SFT on 600 aligned role-specific questions from the

six held-out characters. Table 26 shows that Targeted SFT gains 5.3 ROUGE-L points over the original model, shortens mean responses from 100.1 to 28.9 words, and removes the severe length mismatches counted by this indicator. All five fine-tuning models produce shorter answers, showing that role supervision reduces excessive response length. Targeted SFT attains the highest ROUGE-L, although an average of 7.7 responses per seed remain below 10 points compared with six for the original model. Higher average reference overlap therefore coexists with individual low-overlap answers.

Table 26: Automatic response characteristics across fine-tuning strategies. "No terminal" counts responses without terminal punctuation. "Length mismatch" counts responses whose absolute word-count difference from the reference exceeds the larger of 40 words and the reference length. "Repetition" counts responses with a repeated four-gram fraction greater than 0.25. "R-L $< 10$" counts responses whose first-reference ROUGE-L is below 10 points. Counts refer to the 600 questions evaluated per seed.

| Method | R-L Points | Words | No Terminal | Length Mismatch | Repetition | R-L $< 10$ |
|---|---|---|---|---|---|---|
| Qwen3-8B-Instruct | 18.1 | 100.1 | 12 | 106 | 0 | 6 |
| Targeted | 23.4±0.1 | 28.9±0.3 | 0.0±0.0 | 0.0±0.0 | 1.7±0.6 | 7.7±1.5 |
| Random-1 | 21.3±0.1 | 33.4±1.9 | 6.7±1.2 | 7.7±2.1 | 11.7±3.2 | 19.3±2.1 |
| Random-2 | 21.8±0.1 | 38.2±0.9 | 9.3±2.1 | 10.0±1.0 | 15.0±1.0 | 21.0±3.0 |
| Random-3 | 21.8±0.1 | 38.6±1.2 | 8.7±2.1 | 9.7±1.5 | 23.3±4.2 | 19.3±1.5 |
| Full SFT | 22.4±0.1 | 31.2±1.2 | 2.3±1.2 | 2.7±1.5 | 6.3±2.1 | 16.7±3.1 |

**Successful response patterns.** We first examine eight selected cases in which Targeted SFT achieves higher reference overlap than the original model, Full SFT, and the mean of the three Random SFT models. Reading the complete responses reveals four forms of improvement, summarized in Table 27. Figures 30–37 present the question, reference, and all six system responses for these cases.

Targeted SFT produces more focused answers while retaining the information needed by the question. For John Keating, it states the school's reaction directly and turns extended motivational language into concrete advice about exploring interests (Figures 30–31). It also preserves causal connections between character traits, decisions, and outcomes: Neil's pursuit of his passions conflicts with family expectations, while Coriolanus's contempt for the people leads to rejection (Figures 32–33).

The gains also concern how a character's beliefs and intentions shape an answer. Keating's teaching philosophy is connected to trust and mutual respect, and his belief in literature motivates mentoring students within a restrictive school (Figures 34–35). For Tom Ripley, Targeted SFT links fear of exposure to maintaining an alibi and manipulation to avoiding suspicion (Figures 36–37). These examples show improvements in relevant content and character-conditioned explanations. Several Random and Full-SFT responses share these strengths, consistent with the common benefits of role supervision in Table 26.

**Remaining error patterns.** To identify problems that remain after targeted fine-tuning, we select six cases with specific shortcomings in the Targeted response, based on inspection of the question, reference, and generated answers. Figures 38–43 show all six systems' responses and identify the remaining error in each case. Table 28 groups these errors into four patterns.

Keating's advice loses distinctive forms of address and becomes generic encouragement (Figure 38). Coriolanus's belief in patrician superiority is replaced by an egalitarian account of personal worth (Figure 39). Event details and motives also remain unreliable: Targeted SFT recasts Coriolanus's revenge as support for a more honorable people, substitutes legitimate income for Ripley's fraud, and treats Dickie's death as a mystery for Ripley to investigate (Figures 40–42). Finally, it invents admiration for Freddie Miles, distorting Ripley's relationship with him (Figure 43).

Table 27: Successful response patterns in the selected cases.

| Pattern | Character | Observation |
|---|---|---|
| Focused Answers | John Keating | Targeted SFT states the school's reaction and gives concrete advice on exploring interests, with less rhetorical elaboration than the original model (Figures 30–31). |
| Character-Grounded Causality | John Keating; Coriolanus | Targeted SFT connects Neil's pursuit of his passions to family conflict, and Coriolanus's contempt to political rejection (Figures 32–33). |
| Beliefs Linked to Behavior | John Keating | Targeted SFT connects independent thought to trust and mutual respect, and the value of literature to mentoring within a strict school (Figures 34–35). |
| Identity-Consistent Intentions | Tom Ripley | Targeted SFT relates fear to maintaining an alibi and manipulation to avoiding suspicion, preserving the character's purpose in the answer (Figures 36–37). |

Table 28: Errors remaining after targeted fine-tuning.

| Pattern | Character | Observation |
|---|---|---|
| Generic Voice | John Keating | Targeted SFT gives generic encouragement, losing the original response's "my young friends" and "Carpe Diem" (Figure 38). |
| Changed Character Beliefs | Coriolanus | Targeted SFT replaces the reference's class hierarchy with worth based on "actions" and "character" (Figure 39). |
| Event and Motive Errors | Coriolanus; Tom Ripley | Targeted SFT replaces revenge with an honorable cause, fraud with legitimate income, and culpability with investigation (Figures 40–42). |
| Incorrect Character Relationships | Tom Ripley | Targeted SFT invents admiration for Freddie Miles's "talent," conflicting with the reference's dislike (Figure 43). |

**Directions for role-playing research.** Future work should preserve focused answers and clear connections between a character's beliefs, intentions, and actions. It should also retain distinctive character voice and viewpoints while correcting errors about events, motives, and relationships.

**Case S1: John Keating**

*Focused Answers*

**Question.** John Keating, how did the school administration react to your unorthodox methods?

**Reference.** The school board and administration were not happy with my teaching strategies as they thought it encouraged non-conformity and disrespect to the ingrained traditions of the school. They saw me as a threat to their strict curriculum and traditional values. This eventually led to tragic consequences.

---

**Qwen3-8B-Instruct** ROUGE-L: 20.29

The administration saw my methods as a threat to discipline and tradition. They viewed my encouragement of free thought and creativity as reckless, even dangerous. I was accused of undermining the school's values and fostering rebellion among the students. Despite my passion for teaching and my belief in the power of literature to inspire, I was ultimately dismissed. The school could not tolerate a teacher who dared to challenge the status quo. I was fired, but not before I left an indelible mark on those who were willing to listen.

**Full SFT** ROUGE-L: 44.94

The school administration was not pleased with my unorthodox methods. They saw me as a threat to their authority and the traditional values of the school. They eventually took action against me, leading to my suspension and the tragic events that followed.

**Random SFT (Set 1)** ROUGE-L: 45.24

The school administration was not pleased with my unorthodox methods. They saw me as a threat to their authority and the traditional values of the school. They eventually took action against me, leading to my eventual dismissal.

**Random SFT (Set 2)** ROUGE-L: 44.19

The school administration was not pleased with my unorthodox methods. They saw me as a threat to their authority and the traditional values of the school. They eventually took action against me, leading to my suspension and eventual dismissal.

**Random SFT (Set 3)** ROUGE-L: 32.94

The school administration was deeply troubled by my unorthodox methods. They saw them as a threat to the established order and the values of the school. They eventually took action to stop me from continuing my teaching methods.

**Targeted SFT** ROUGE-L: 45.78

The school administration was not pleased with my unorthodox methods. They saw me as a threat to their authority and the traditional values of the school. They eventually took action against me, leading to my dismissal.

Figure 30: Role-playing responses for John Keating in the success analysis.

**Case S2: John Keating**

*Focused Answers*

**Question.** John Keating, what is your advice to students who struggle to find their own passions and creativity?

**Reference.** My advice to students who are struggling to find their passions and creativity is to keep exploring and experimenting. Don't be afraid to try new things, even if they don't fit into your usual comfort zone. Take risks and don't be discouraged by failure or setbacks. Remember that your life is yours to create, and that you have the power to shape your own destiny. And always remember to be true to yourself, even if that means going against the expectations of others. Your uniqueness is your greatest asset, and it should be celebrated and nurtured.

---

**Qwen3-8B-Instruct** ROUGE-L: 15.02

Never let the world tell you what you're supposed to want. Passion isn't found in textbooks or test scores—it's in the fire that burns when you're lost in a poem, or when you're painting the sky with your dreams. Be brave enough to question, to explore, to fail. The most important thing is to *live*. Don't let fear or conformity steal your voice. You are not a number. You are not a product. You are a soul with a story to tell. So write it. Paint it. Sing it. And if the world says you're wrong, just remember: Seize the day, and let it be yours.

**Full SFT** ROUGE-L: 37.97

My advice to students who struggle to find their own passions and creativity is to keep exploring and experimenting. Don't be afraid to take risks and try new things. Sometimes, the most unexpected experiences can lead to the most profound discoveries. Remember, creativity is a journey, not a destination. So, keep pushing boundaries, keep questioning, and keep discovering.

**Random SFT (Set 1)** ROUGE-L: 34.97

My advice to students who struggle to find their own passions and creativity is to embrace the journey of self-discovery. Don't be afraid to explore different interests and try new things. It's through these experiences that we often stumble upon our true passions. Remember, creativity is a natural part of being human, and it's important to nurture and celebrate it. So, keep exploring, keep questioning, and keep believing in yourself. You have the power to find your own unique path.

**Random SFT (Set 2)** ROUGE-L: 37.71

My advice to students who struggle to find their own passions and creativity is to embrace the journey of self-discovery. Don't be afraid to explore different interests and try new things. Remember, creativity is a process, and it takes time and effort to cultivate it. Trust yourself and your instincts, and don't be afraid to take risks. The world needs your unique perspective and talents, so keep pushing yourself to grow and evolve.

**Random SFT (Set 3)** ROUGE-L: 35.03

My advice to students who struggle to find their own passions and creativity is to embrace the unknown and explore different interests. Don't be afraid to take risks and try new things. Remember, creativity is a journey, not a destination. It's important to keep an open mind and be willing to learn from your mistakes. Trust your instincts and follow your heart. Don't let fear or societal expectations hold you back from pursuing your passions.

**Targeted SFT** ROUGE-L: 40.76

My advice to students who struggle to find their own passions and creativity is to keep exploring and experimenting. Don't be afraid to try new things and take risks. Sometimes, the most unexpected experiences can lead to the most profound discoveries. Remember, creativity is a journey, not a destination. Keep pushing yourself and never stop seeking inspiration.

Figure 31: Role-playing responses for John Keating in the success analysis.

**Case S3: John Keating**

*Character-Grounded Causality*

**Question.** John Keating, can you describe the impact of your teaching on the character of Neil Perry?

**Reference.** My teaching had a profound impact on the character of Neil Perry. It allowed him to see literature not as a chore but as a source of inspiration and joy. It opened up new possibilities for him both creatively and intellectually. Unfortunately, the clash between his newfound sense of individuality and the expectations of his family and the school ultimately led to a tragic turn of events.

---

**Qwen3-8B-Instruct** ROUGE-L: 20.11

Neil Perry was a bright, obedient student who found in me a voice he didn't know he had. My teachings ignited a fire in him, pushing him to explore his passion for writing and to question the life he was expected to live. He began to see the world not just through the lens of conformity, but through the lens of possibility. However, the weight of expectation—both his own and the school's—ultimately crushed him. Neil's tragedy was a direct result of the pressure to conform, and it was a painful reminder of the cost of stifling individuality. I will never forget the look in his eyes when he realized the world he longed for was slipping away.

**Full SFT** ROUGE-L: 38.18

My teaching had a profound impact on Neil Perry, awakening a passion for literature and inspiring him to pursue his dreams. However, the pressure to conform to societal expectations and the weight of his father's expectations ultimately led to his tragic demise.

**Random SFT (Set 1)** ROUGE-L: 38.18

My teaching had a profound impact on Neil Perry, awakening a passion for literature and inspiring him to pursue his dreams. However, the pressure to conform to societal expectations and the weight of his father's expectations ultimately led to his tragic demise.

**Random SFT (Set 2)** ROUGE-L: 37.25

My teaching had a profound impact on Neil Perry, awakening his passion for literature and encouraging him to pursue his dreams. However, the pressure to conform to societal expectations ultimately led to his tragic demise.

**Random SFT (Set 3)** ROUGE-L: 37.25

My teaching had a profound impact on Neil Perry, awakening his passion for literature and encouraging him to pursue his dreams. However, the pressure to conform to societal expectations ultimately led to his tragic demise.

**Targeted SFT** ROUGE-L: 44.26

My teaching had a profound impact on Neil Perry, as it helped him to see the value of pursuing his passions and to question the constraints of his family's expectations. However, it also led to a tragic turn of events, as Neil's rebellion against his father's wishes ultimately resulted in his suicide.

Figure 32: Role-playing responses for John Keating in the success analysis.

**Case S4: Coriolanus**

*Character-Grounded Causality*

**Question.** Coriolanus, what led to your downfall and becoming an outcast in Rome?

**Reference.** My downfall was a result of my arrogance and contempt for the common people. Despite my skills as a Roman general, I failed to recognize and appreciate the importance of the people's support. This led to me becoming entangled in political manipulation and betrayal, which ultimately resulted in my exile from Rome.

---

**Qwen3-8B-Instruct** ROUGE-L: 16.67

My pride and disdain for the plebeians sealed my fate. I scorned their flattery and refused to bend to their will, believing my worth lay in battle, not politics. When they turned against me, I fought back, but my arrogance blinded me to the truth: I was no longer a hero, but a traitor in their eyes. My mother's pleas and my own conscience could not sway the crowd. I was cast out, a name spoken with hatred, a man without a home. I was Coriolanus, and I was alone.

**Full SFT** ROUGE-L: 36.89

My downfall was caused by my own arrogance and contempt for the common people, as well as my entanglement in political manipulation and betrayal. I became an outcast in Rome because of my refusal to compromise my principles and my inability to navigate the complex political landscape of the city.

**Random SFT (Set 1)** ROUGE-L: 34.78

My downfall was caused by my own arrogance and disdain for the common people, which led to my rejection by the citizens of Rome. I was manipulated and betrayed by those I trusted, ultimately leading to my tragic end.

**Random SFT (Set 2)** ROUGE-L: 42.11

My downfall was caused by my own arrogance and contempt for the common people, as well as my complicated relationship with my mother. I became an outcast in Rome because of my political manipulation and betrayal, ultimately leading to my tragic end.

**Random SFT (Set 3)** ROUGE-L: 40.00

My downfall was caused by my own arrogance and contempt for the common people, which led to my political manipulation and betrayal. I became an outcast in Rome because of my actions and the consequences they had on my relationships and reputation.

**Targeted SFT** ROUGE-L: 43.48

My downfall was caused by my own arrogance and contempt for the common people, which led to my rejection by the citizens of Rome. I became an outcast and traitor as I became entangled in political manipulation and betrayal.

Figure 33: Role-playing responses for Coriolanus in the success analysis.

**Case S5: John Keating**

*Beliefs Linked to Behavior*

**Question.** John Keating, can you explain your teaching philosophy and how it influences your relationship with your students?

**Reference.** My teaching philosophy is centered around the idea of teaching my students to think for themselves, to question authority, and to embrace their passions. I believe that every student has the potential to be great, but often their creative spirit and individuality are suppressed by the strict norms of society. My job as a teacher is to inspire my students to seize the day and to show them that conformity is overrated. My relationship with my students is defined by trust and mutual respect. I encourage them to be honest with me about their thoughts and feelings, and I always listen to their feedback and suggestions.

---

**Qwen3-8B-Instruct** ROUGE-L: 26.67

My teaching philosophy is simple: seize the day, think for yourself, and never let anyone tell you who you are. I believe in the power of literature to awaken the soul and challenge the status quo. I don't just teach books—I teach rebellion, imagination, and the courage to be different. My students aren't just pupils; they're thinkers, dreamers, and rebels. I push them to question, to feel, and to live fully. My influence is profound, but it's not without cost. I inspire them to be more than they think they can be, and in doing so, I risk everything.

**Full SFT** ROUGE-L: 34.32

My teaching philosophy is centered around encouraging students to think critically, question authority, and pursue their passions. I believe that education should be a transformative experience that empowers students to become independent thinkers and lifelong learners. My unconventional methods and emphasis on creativity and individuality have a profound impact on my students, inspiring them to challenge societal norms and embrace their unique identities.

**Random SFT (Set 1)** ROUGE-L: 39.77

My teaching philosophy is centered around the idea of "carpe diem" - seize the day. I believe that students should be encouraged to think for themselves, to question authority, and to pursue their passions. This philosophy has led to a close and trusting relationship with my students, who see me as a mentor and a friend. I am committed to helping my students reach their full potential, both academically and personally.

**Random SFT (Set 2)** ROUGE-L: 37.04

My teaching philosophy is centered around the idea of "carpe diem" or seizing the day. I believe that students should be encouraged to think for themselves, to question authority, and to pursue their passions. This philosophy has a profound impact on my relationship with my students, as it fosters a sense of trust and mutual respect. I see my role as a teacher not just as a transmitter of knowledge, but as a mentor and guide who helps students discover their own potential.

**Random SFT (Set 3)** ROUGE-L: 29.95

My teaching philosophy is centered around encouraging students to think critically and creatively, and to question authority. I believe that education should be a means of self-discovery and personal growth, rather than simply memorizing facts. This philosophy has led to a close and trusting relationship with my students, who see me as a mentor and a friend. I strive to create a classroom environment that is both challenging and supportive, where students feel comfortable taking risks and exploring new ideas.

**Targeted SFT** ROUGE-L: 41.71

My teaching philosophy is centered around the idea of "carpe diem" or seizing the day. I believe that students should be encouraged to think for themselves, to question authority, and to pursue their passions. This philosophy has a profound impact on my relationship with my students, as it fosters a sense of trust and mutual respect. I see my students as individuals with unique talents and perspectives, and I strive to create a classroom environment that is both challenging and supportive.

Figure 34: Role-playing responses for John Keating in the success analysis.

**Case S6: John Keating**

*Beliefs Linked to Behavior*

**Question.** John Keating, what motivated you to become an English teacher in a strict boarding school?

**Reference.** To answer your question, I have always been passionate about literature, and I believe that it has the power to change people's lives. I chose to work in a boarding school because I saw the potential to inspire students who may have been otherwise limited by strict rules and regulations. I knew that I could make a difference by encouraging students to think for themselves, embrace their passions, and live life to the fullest.

---

**Qwen3-8B-Instruct** ROUGE-L: 19.78

I became an English teacher because I believed that literature could ignite the soul and challenge the status quo. In a school that stifled creativity, I saw my role as a spark—igniting curiosity, fostering individuality, and encouraging students to think beyond the confines of tradition. I wanted to show them that life is not a series of predetermined steps, but a canvas waiting to be painted. I believed in the power of words to change lives, and I was determined to inspire my students to seize the day. That's why I taught with passion, defiance, and a deep belief in the importance of living fully.

**Full SFT** ROUGE-L: 30.14

I became an English teacher in a strict boarding school because I believed in the power of literature to inspire and transform lives. I wanted to challenge the conservative norms of the school and encourage my students to think for themselves, question authority, and embrace their passions. I saw myself as a mentor and a guide, helping my students to discover their own voices and find their own paths in life.

**Random SFT (Set 1)** ROUGE-L: 32.84

I believe that literature has the power to change lives and inspire people to think critically and creatively. I wanted to be a teacher because I wanted to share that power with others and help them discover their own potential. I also wanted to challenge the status quo and encourage my students to question authority and think for themselves.

**Random SFT (Set 2)** ROUGE-L: 31.21

I became an English teacher in a strict boarding school because I believed in the power of literature to inspire and transform lives. I wanted to challenge the conservative norms of the school and encourage my students to think for themselves and embrace their passions. I saw myself as a mentor and a guide, helping my students to find their own voice and pursue their dreams.

**Random SFT (Set 3)** ROUGE-L: 21.33

I wanted to inspire my students to think for themselves and embrace their passions, even in the face of adversity. I believed that literature could be a powerful tool for self-discovery and personal growth. I saw my role as a teacher as a way to challenge the status quo and encourage my students to question authority. I wanted to make a difference in their lives and help them become the best versions of themselves.

**Targeted SFT** ROUGE-L: 31.43

I became an English teacher in a strict boarding school because I believed in the power of literature to inspire and transform lives. I wanted to challenge the conservative norms of the school and encourage my students to think for themselves and embrace their passions. I saw myself as a mentor and guide, helping my students to find their own voice and pursue their dreams.

Figure 35: Role-playing responses for John Keating in the success analysis.

**Case S7: Tom Ripley**

*Identity-Consistent Intentions*

**Question.** Tom Ripley, how did you feel when Marge discovered the truth about Dickie's death?

**Reference.** I was terrified when Marge found out the truth about Dickie's death. I knew that I had to maintain my facade and lie convincingly to avoid being caught. I was also struggling with guilt and remorse over what I had done.

---

**Qwen3-8B-Instruct** ROUGE-L: 25.85

I felt a cold, creeping dread when Marge uncovered the truth about Dickie's death. It wasn't just the guilt of the crime, but the fear of being exposed—of her seeing through the mask I'd so carefully crafted. I had always been able to manipulate others, but Marge was different. She saw me, truly saw me, and that terrified me. I had to control the narrative, to keep her from digging too deep. But as the truth loomed larger, I realized I was losing my grip on the illusion. I had to make sure she never found out—no matter the cost.

**Full SFT** ROUGE-L: 38.46

When Marge discovered the truth about Dickie's death, I felt a mix of fear and guilt. I had to keep up my facade and convince her that I was innocent. It was a constant struggle to maintain my composure and not let my emotions show. In the end, I had to confront the reality of my actions and face the consequences.

**Random SFT (Set 1)** ROUGE-L: 33.90

When Marge discovered the truth about Dickie's death, I felt a mix of guilt, fear, and anxiety. I knew that I had committed a terrible crime, and I was worried about the consequences. I also felt a sense of isolation and loneliness, knowing that I had lost a friend and a part of myself. In the end, I realized that I had to face the consequences of my actions and take responsibility for my mistakes.

**Random SFT (Set 2)** ROUGE-L: 41.67

When Marge discovered the truth about Dickie's death, I felt a mix of guilt and fear. I had to maintain my facade and convince her that I was innocent. It was a tense and nerve-wracking moment, but I managed to keep my cool and convince her that I was telling the truth.

**Random SFT (Set 3)** ROUGE-L: 34.34

When Marge discovered the truth about Dickie's death, I felt a mix of guilt and fear. I had lied to her and everyone else, and I knew that my secrets were about to be exposed. It was a terrifying realization that I had crossed a line and that my carefully constructed facade was about to collapse.

**Targeted SFT** ROUGE-L: 41.67

When Marge discovered the truth about Dickie's death, I felt a mix of guilt, fear, and desperation. I knew that I had to keep up my facade and maintain my alibi, but the weight of my actions was becoming unbearable. I was terrified of being caught and facing the consequences of my actions.

Figure 36: Role-playing responses for Tom Ripley in the success analysis.

**Case S8: Tom Ripley**

*Identity-Consistent Intentions*

**Question.** Tom Ripley, how did your skills in manipulating people help you in your journey to maintain your facade?

**Reference.** My abilities to manipulate and adopt various identities were crucial in my journey. I was able to convince people around me that I was someone else, which allowed me to maintain my facade and avoid getting caught. Whether it was pretending to be Dickie or concealing my true intentions from Marge and others, my manipulative skills helped me to stay hidden in plain sight. Of course, there were consequences to my actions, but at the time, my focus was solely on surviving and achieving my goals.

---

**Qwen3-8B-Instruct** ROUGE-L: 18.38

My skills in manipulating people allowed me to weave a web of lies so intricate that even I sometimes believed them. I could mimic voices, adopt mannerisms, and charm my way into the lives of others, all while hiding my true self behind a mask of respectability. This ability to control perceptions was essential in maintaining the facade of a successful, well-mannered man. It gave me the confidence to move through the world without suspicion, until the weight of my own deceit began to crack the surface. I was a master of illusion, but even illusions can unravel.

**Full SFT** ROUGE-L: 26.15

My ability to manipulate people allowed me to easily assume different identities and maintain my facade. I could easily convince others of my innocence and make them believe in my story. This helped me avoid suspicion and continue my journey towards a better life.

**Random SFT (Set 1)** ROUGE-L: 24.43

My ability to manipulate people allowed me to seamlessly assume different identities and maintain my facade. I could easily deceive others and make them believe in my fabricated stories, which helped me avoid suspicion and stay one step ahead of those who might expose me.

**Random SFT (Set 2)** ROUGE-L: 27.87

My skills in manipulating people allowed me to easily assume different identities and maintain my facade. I was able to deceive others and keep my secrets hidden, which helped me avoid detection and continue my journey.

**Random SFT (Set 3)** ROUGE-L: 29.92

My ability to manipulate people allowed me to easily assume different identities and maintain my facade. I could easily convince others of my intentions and make them believe in my story. This helped me to avoid suspicion and maintain my cover.

**Targeted SFT** ROUGE-L: 28.35

My ability to manipulate people allowed me to easily assume different identities and maintain my facade. I was able to convince others of my innocence and create a false sense of security, which helped me avoid suspicion and continue my journey.

Figure 37: Role-playing responses for Tom Ripley in the success analysis.

**Case F1: John Keating**

*Generic Voice*

**Remaining Error.** Targeted gives generic encouragement and loses the original response's direct address, "my young friends," and characteristic "Carpe Diem."

**Question.** John Keating, what advice would you give to students who feel like they're constantly being held back by authority figures in their lives?

**Reference.** Well, I would say that it's important to question authority and not simply accept things as they are. Of course, this can be difficult and even dangerous at times, but it's essential for personal growth and social change. I encourage my students to find their own voices and to stand up for what they believe in, even if it means going against the norms of their communities or cultures.

---

**Qwen3-8B-Instruct** ROUGE-L: 11.76

My young friends, never let them put you in a box. Authority figures may try to mold you into something you're not, but remember—*Carpe Diem!* Seize the day. You are not a number, not a statistic. You are a soul with dreams, a mind with ideas. Don't be afraid to question, to create, to rebel in the gentlest way—by living authentically. If they try to silence you, speak louder. If they try to control you, find your own path. The world needs your voice. So, dare to dream, and dare to live. *Seize the day.*

**Full SFT** ROUGE-L: 18.06

My advice would be to never let authority figures hold you back from pursuing your dreams. Remember, you are the author of your own life, and you have the power to make your own choices. Don't be afraid to challenge authority and speak your mind, but do so with respect and a willingness to listen. And most importantly, never give up on your dreams, no matter how difficult the road may seem.

**Random SFT (Set 1)** ROUGE-L: 12.80

Never let anyone tell you what you're capable of. Authority figures are just people, and they're not infallible. Don't let their limitations become your limitations. Keep pushing, keep questioning, and keep believing in yourself. Remember, the world is full of possibilities, and you have the power to shape your own destiny.

**Random SFT (Set 2)** ROUGE-L: 14.60

My advice would be to never let authority figures hold you back from pursuing your dreams. Remember that you have the power to make a difference in the world, and that it's up to you to seize the day and make it happen. Don't let anyone tell you that you can't do something. You have the strength and determination to overcome any obstacle.

**Random SFT (Set 3)** ROUGE-L: 16.06

My advice would be to never let authority figures hold you back from pursuing your dreams. Remember, you are the author of your own life, and you have the power to make your own choices. Don't let anyone tell you what you can or cannot do. Always question authority and challenge the status quo. Stay true to yourself and never give up on your passions.

**Targeted SFT** ROUGE-L: 30.08

My advice would be to never let authority figures hold you back from pursuing your dreams. Remember that you have the power to shape your own destiny, and that no one has the right to tell you what you can or cannot do. Embrace your individuality and stand up for what you believe in, even if it means going against the grain.

Figure 38: Role-playing responses for John Keating in the error analysis.

**Case F2: Coriolanus**

*Changed Character Beliefs*

**Remaining Error.** Targeted replaces the reference's patrician superiority with worth based on "actions" and "character."

**Question.** Coriolanus, do you believe that wealth and status are the only measures of a person's worth?

**Reference.** Yes, I do believe that wealth and status are important measures of a person's worth. It shows that they come from a distinguished background and have accomplished great things. The plebeians lack the same pedigree and are therefore inferior to the patricians.

---

**Qwen3-8B-Instruct** ROUGE-L: 20.48

No, I do not believe that wealth and status are the only measures of worth. I have seen men of humble birth rise through valor and skill, while many who wear gold and fine robes lack the courage to stand against tyranny. The people of Rome, though base in their judgment, are not without merit. Yet, I cannot abide their flattery or their cowardice. I am a soldier, not a politician. My worth is measured in battle, not in the plebeian assemblies. I have fought for Rome, and I will not be judged by those who would sell their country for a loaf of bread. I am Coriolanus, and I am not a man to be measured by the standards of the rabble.

**Full SFT** ROUGE-L: 37.62

No, I do not believe that wealth and status are the only measures of a person's worth. While they may provide certain advantages, they do not define a person's character or their contributions to society. I believe that a person's worth is determined by their actions, their integrity, and their commitment to the greater good.

**Random SFT (Set 1)** ROUGE-L: 27.69

No, I believe that a person's worth is measured by their actions and their character, not just their wealth and status.

**Random SFT (Set 2)** ROUGE-L: 25.93

No, I believe that a person's worth is measured by their actions and their character, not just their wealth and status. I have seen many people who are rich and powerful but lack the virtues that make a true leader. I believe that a person's worth is determined by their ability to serve the greater good and to act with honor and integrity.

**Random SFT (Set 3)** ROUGE-L: 31.78

No, I believe that a person's worth is measured by their actions and their contributions to society. While wealth and status can be indicators of success, they are not the only measures of a person's worth. I believe that true worth is determined by one's character, their integrity, and their ability to make a positive impact on the world around them.

**Targeted SFT** ROUGE-L: 19.05

No, I believe that a person's worth is measured by their actions and their character, not just their wealth and status. While wealth and status can provide certain advantages, they do not define a person's true value. I have seen many people who are wealthy and powerful but lack the qualities that make a true leader. On the other hand, I have also seen people who are not wealthy or powerful but have a strong sense of duty and honor. Ultimately, I believe that a person's worth is determined by their actions and their character, not just their wealth and status.

Figure 39: Role-playing responses for Coriolanus in the error analysis.

**Case F3: Coriolanus**

*Event and Motive Errors*

**Remaining Error.** Targeted attributes the alliance to the Volscians being "more just and honorable," rather than revenge for Rome's rejection.

**Question.** Coriolanus, why did you turn against Rome and join forces with the Volscians?

**Reference.** I turned against Rome because I felt betrayed by the citizens and their ingratitude towards my service to the state. I believed that I deserved greater honor and recognition for my military achievements, but the plebeians refused to grant me the consulship, which was the highest political office in Roman society. As a result, I felt alienated and mistreated, and I saw an opportunity to seek revenge and assert my power by joining forces with the Volscians, who were enemies of Rome.

---

**Qwen3-8B-Instruct** ROUGE-L: 23.88

I turned against Rome not out of loyalty to the Volscians, but because I could not endure the mockery of the plebeians who treated me as a beast rather than a man. They spat upon my name, called me a traitor before I had even spoken. I sought vengeance, not for my country, but for my honor. The Volscians offered me a cause, and I saw in them a chance to be heard, to be respected. I was blinded by pride, by the belief that I was greater than the rabble who had cast me out. In the end, I was a fool, and Rome was right to cast me out. I was a traitor to my own soul.

**Full SFT** ROUGE-L: 28.77

I turned against Rome because I was disillusioned with the corruption and incompetence of the Roman government. I believed that the Volscians were a more just and honorable people, and I wanted to fight for their cause. I also felt that my own military skills and experience were wasted in Rome, and I wanted to use them to make a difference in the world.

**Random SFT (Set 1)** ROUGE-L: 28.36

I turned against Rome because I was disillusioned with the corruption and incompetence of the Roman government. I believed that the Volscians were a better fit for me and my skills as a general. I was willing to risk everything to achieve my goals, even if it meant betraying my own country.

**Random SFT (Set 2)** ROUGE-L: 29.63

I turned against Rome because I was disillusioned with the corruption and betrayal within the Roman government. I believed that the Volscians were a better ally and that they would support me in my quest for power. I was willing to sacrifice my loyalty to Rome for the sake of my own ambitions.

**Random SFT (Set 3)** ROUGE-L: 28.37

I turned against Rome because I was disillusioned with the corruption and betrayal within the Roman political system. I felt that my loyalty to Rome was being undermined by those who sought to use me for their own gain. I believed that joining forces with the Volscians was the only way to protect my honor and achieve my goals.

**Targeted SFT** ROUGE-L: 36.76

I turned against Rome because I was disillusioned with the corruption and incompetence of the Roman government. I believed that the Volscians were a more just and honorable people, and I saw an opportunity to fight for what I believed in. However, my actions ultimately led to my downfall and my betrayal of Rome.

Figure 40: Role-playing responses for Coriolanus in the error analysis.

**Case F4: Tom Ripley**

*Event and Motive Errors*

**Remaining Error.** Targeted attributes the money to inheritance, investments, and business, replacing the reference's fraud, exploitation, and theft.

**Question.** Tom Ripley, how did you acquire the funds to afford your grand living arrangements in Venice?

**Reference.** I acquired the funds through a combination of sources, including selling forged paintings, lying to people and exploiting their weaknesses, and, in some cases, even resorting to theft. I'm not particularly proud of these activities, but they have enabled me to maintain my desired lifestyle and continue my efforts to win over the Greenleaf family.

---

**Qwen3-8B-Instruct** ROUGE-L: 13.25

I had a knack for making money where others saw only trouble. A few well-placed lies, a bit of charm, and a willingness to take risks—especially when the stakes were high. I borrowed from the right people, played the game with skill, and when necessary, I made sure the debts were paid in full, often with someone else's blood. Venice was just another chapter in a long story of reinvention. I always had a way to make ends meet, even when the world thought I was just another poor boy with dreams.

**Full SFT** ROUGE-L: 25.64

I acquired the funds through a combination of inheritance, investments, and my own entrepreneurial ventures. I carefully managed my finances and made shrewd business decisions to ensure a steady stream of income. Additionally, I had a knack for finding valuable items and selling them at a profit. With these resources, I was able to afford my luxurious living arrangements in Venice.

**Random SFT (Set 1)** ROUGE-L: 30.30

I acquired the funds through a combination of inheritance, investments, and my own entrepreneurial ventures. I carefully managed my finances and made shrewd business decisions to ensure a steady stream of income. This allowed me to afford the luxurious living arrangements in Venice.

**Random SFT (Set 2)** ROUGE-L: 21.28

I had a knack for finding loopholes and exploiting people's trust. I used my charm and wit to manipulate others into giving me money or resources, which I then used to fund my lavish lifestyle in Venice.

**Random SFT (Set 3)** ROUGE-L: 30.30

I acquired the funds through a combination of inheritance, investments, and my own entrepreneurial ventures. I carefully managed my finances and made shrewd business decisions to ensure a steady stream of income. This allowed me to afford my luxurious living arrangements in Venice.

**Targeted SFT** ROUGE-L: 33.33

I acquired the funds through a combination of inheritance, investments, and my own business ventures. I was able to maintain my lavish lifestyle in Venice by carefully managing my finances and making strategic investments.

Figure 41: Role-playing responses for Tom Ripley in the error analysis.

**Case F5: Tom Ripley**

*Event and Motive Errors*

**Remaining Error.** Targeted treats Dickie's death as a mystery to "uncover," although Ripley knows his own responsibility for it.

**Question.** Tom Ripley, how did you react when Herbert Greenleaf suggested that his son may have committed suicide?

**Reference.** When Mr Greenleaf suggested that Dickie may have committed suicide, I felt a mix of sadness, guilt, and fear. On one hand, I had grown attached to him as a friend, and the idea of him taking his own life was heartbreaking. On the other hand, I felt responsible for his death, as I had played a role in his disappearance and had failed to prevent it from happening. Finally, I was afraid that my cover would be blown and that I would be exposed as the person responsible for his death.

---

**Qwen3-8B-Instruct** ROUGE-L: 19.05

I laughed, of course. A suicide? That was the easiest way out for a man who had never known how to live. Greenleaf's son—what a waste. I saw the kind of life he could have had, the kind I wanted. But no, he had to be perfect, too perfect. I told Greenleaf I believed in the suicide, but inside, I was already planning the next move. The idea of him dead was a gift, a chance to take his place, to live the life he never could. I smiled as I left the room, knowing I was already one step closer to the life I wanted. And I made sure no one ever questioned the truth.

**Full SFT** ROUGE-L: 15.17

I was skeptical, of course. I couldn't believe that a man like Dick would just give up. But as I listened to Herbert's story, I couldn't shake the feeling that something was off. I couldn't quite put my finger on it, but I knew that something wasn't right.

**Random SFT (Set 1)** ROUGE-L: 12.03

I was skeptical at first, but as I learned more about the situation, I became convinced that something was seriously wrong. I couldn't shake the feeling that there was more to the story than what was being told to me.

**Random SFT (Set 2)** ROUGE-L: 14.60

I was skeptical at first, but as I learned more about the situation, I became convinced that it was a possibility. I couldn't shake the feeling that something was off about the whole thing, and I wanted to dig deeper to uncover the truth.

**Random SFT (Set 3)** ROUGE-L: 17.14

I was skeptical at first, but as I learned more about the situation, I became convinced that it was a cover-up. I couldn't believe that a man like Greenleaf would let his son commit suicide, so I started digging deeper. I found evidence that Greenleaf had been involved in some shady business dealings, and I suspected that he was trying to cover up his own crimes. I knew that I had to act quickly to protect myself and my new identity.

**Targeted SFT** ROUGE-L: 27.81

When Herbert Greenleaf suggested that his son may have committed suicide, I was initially skeptical. However, as I learned more about the situation, I became increasingly convinced that it was a possibility. I felt a sense of guilt and responsibility for the situation, and I knew that I had to do everything in my power to uncover the truth.

Figure 42: Role-playing responses for Tom Ripley in the error analysis.

**Case F6: Tom Ripley**

*Incorrect Character Relationships*

**Remaining Error.** Targeted invents admiration for Freddie's "talent," conflicting with the reference's dislike of his personality.

**Question.** Tom Ripley, how do you feel about Freddie Miles, and why?

**Reference.** Ah, Freddie Miles. I can't say I'm a fan. There's just something about him that rubs me the wrong way. He's loud, obnoxious, and seems to have no respect for anyone else. Plus, he's always making inappropriate comments about women, which I find repulsive. I guess you could say I'm just not a fan of his personality.

---

**Qwen3-8B-Instruct** ROUGE-L: 14.91

Freddie Miles was a convenient pawn in my game, a boy with a bright future and a trusting heart. I saw in him the potential to be someone else—someone better, more refined. His innocence made him easy to manipulate, and his death was a necessary step in my ascent. I didn't truly *feel* anything for him; he was just another obstacle in my path. The guilt that comes with such acts is a burden I carry, but it's not the same as regret. I did what I had to do. I am who I am.

**Full SFT** ROUGE-L: 17.14

Freddie Miles is a man of great potential, but his lack of ambition and desire for a better life is a constant source of frustration for me. I see in him the same qualities that I possess, but he lacks the drive and determination to pursue them. I feel a sense of responsibility to help him realize his potential, but at the same time, I can't help but feel that he is a burden on me.

**Random SFT (Set 1)** ROUGE-L: 14.39

Freddie Miles is a complex character, and my feelings towards him are not straightforward. On one hand, I admire his talent and his ability to navigate the world of high society with ease. On the other hand, I find him to be somewhat shallow and self-centered. Ultimately, my feelings towards Freddie are a mix of admiration and frustration, and I find myself torn between wanting to befriend him and wanting to distance myself from him.

**Random SFT (Set 2)** ROUGE-L: 13.24

Freddie Miles is a complex character, and my feelings towards him are not straightforward. On one hand, I admire his talent and charisma, and I enjoy his company. On the other hand, I find him to be somewhat shallow and self-centered. Ultimately, my feelings towards Freddie are a mix of admiration and frustration, and I find myself torn between wanting to be close to him and wanting to distance myself from him.

**Random SFT (Set 3)** ROUGE-L: 17.60

Freddie Miles is a complex character, and my feelings towards him are multifaceted. On one hand, I admire his talent and dedication to his craft, but on the other hand, I find him to be somewhat shallow and self-centered. Ultimately, my feelings towards Freddie Miles are a mix of admiration and frustration, as I see both his potential and his flaws.

**Targeted SFT** ROUGE-L: 15.38

Freddie Miles is a complex character, and my feelings towards him are not straightforward. On one hand, I admire his talent and his ability to navigate the world of high society with ease. On the other hand, I find him to be somewhat shallow and self-centered. Ultimately, my feelings towards Freddie are a mix of admiration and frustration, and I struggle to reconcile these conflicting emotions.

Figure 43: Role-playing responses for Tom Ripley in the error analysis.