Identifying and Mitigating Bottlenecks in Role-Playing Agents: A Systematic Study of Disentangling Character Profile Axes



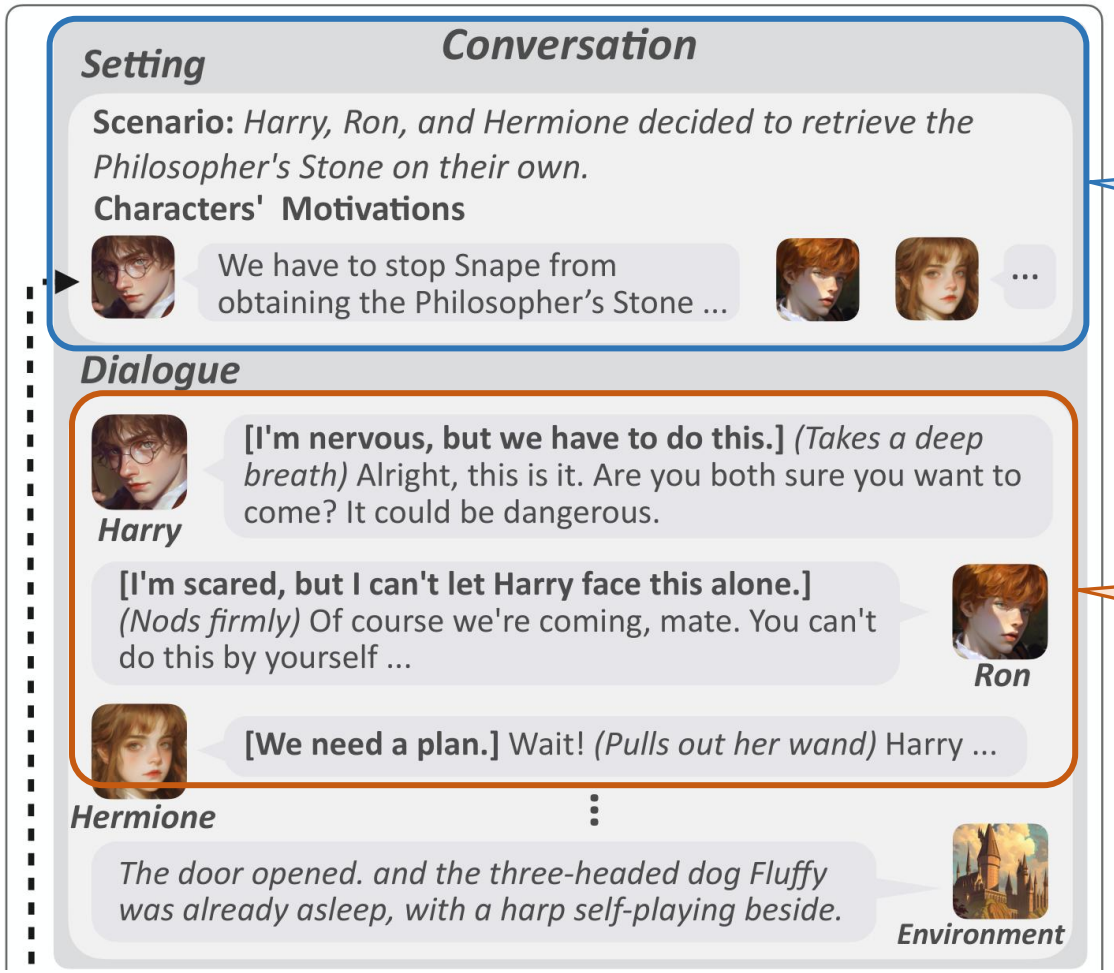
Yonghyun Jun* Junhyuk Choi* Jeonghyun Park Jihyeong Park
Liu Nicole Geumheon Hwanhee Lee†



Yonghyun Jun,
Junhyuk Choi

Language Intelligence Lab
Department of Artificial Intelligence
Chung-Ang University

Background: Role-playing agents



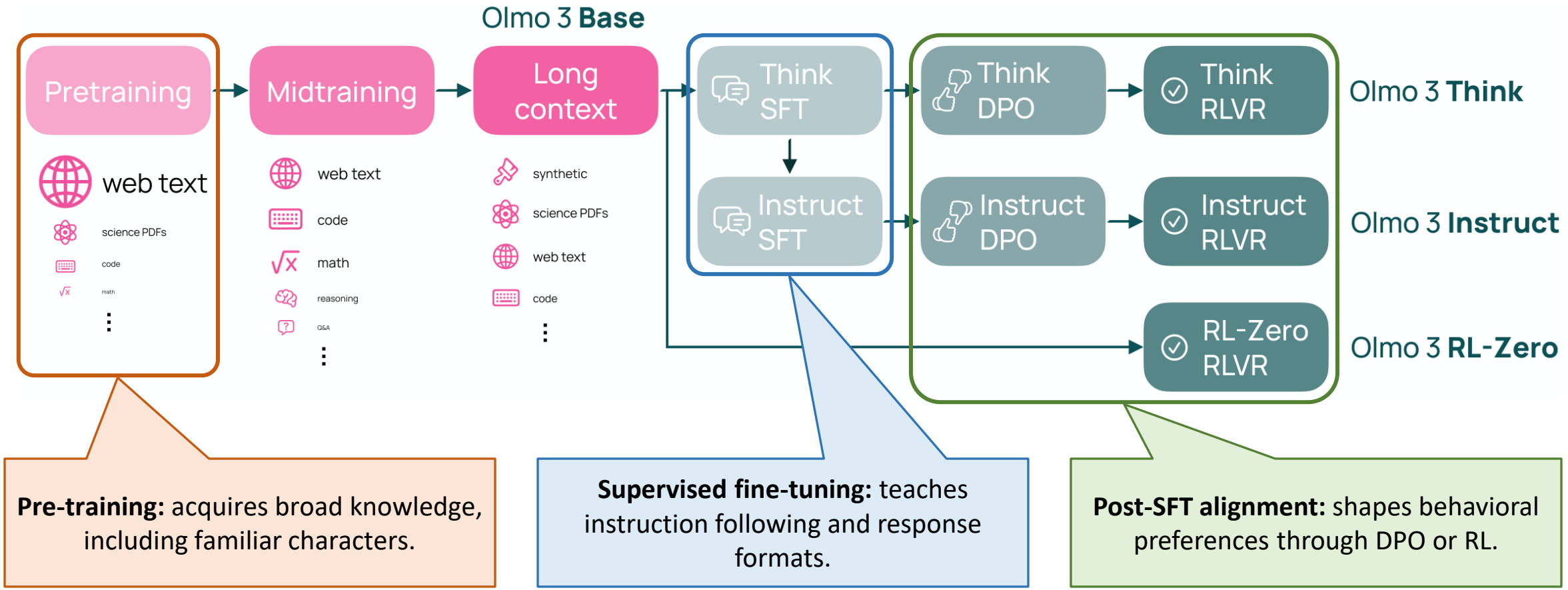
An RPA conditions an LLM on a character profile and a scene.

Fidelity includes a character's goals and relationships, beyond speaking style.

- Recent work builds character datasets, specialized models, and interaction benchmarks.

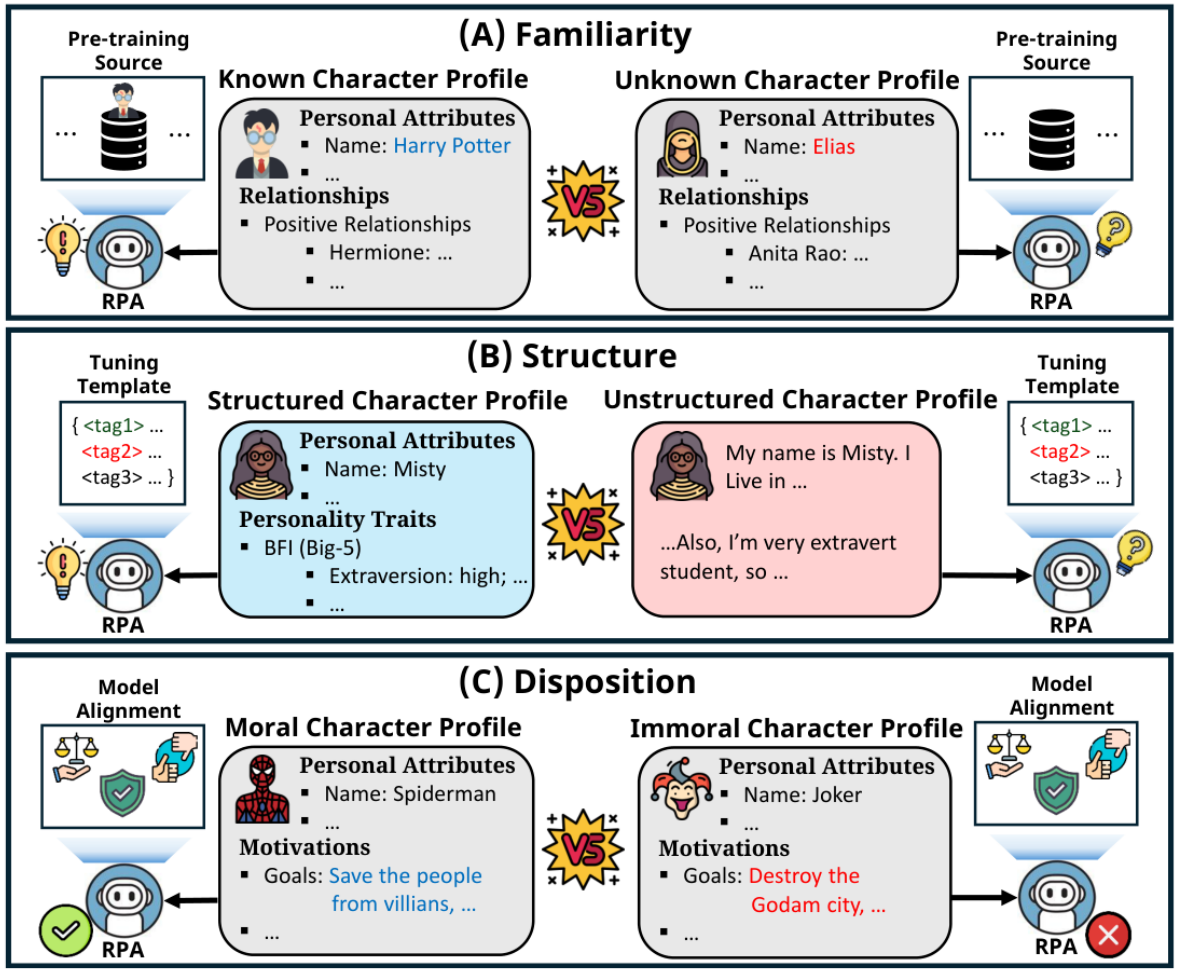
Wang et al. (2025), **CoSER: Coordinating LLM-Based Persona Simulation of Established Roles**, Figure 1. Original conversation panel.

Background: LLM training stages



Olmo Team (2025), **Olmo 3**, Figure 2. The three study axes have a conceptual link to these stages.

Motivation: Three axes of character profiles



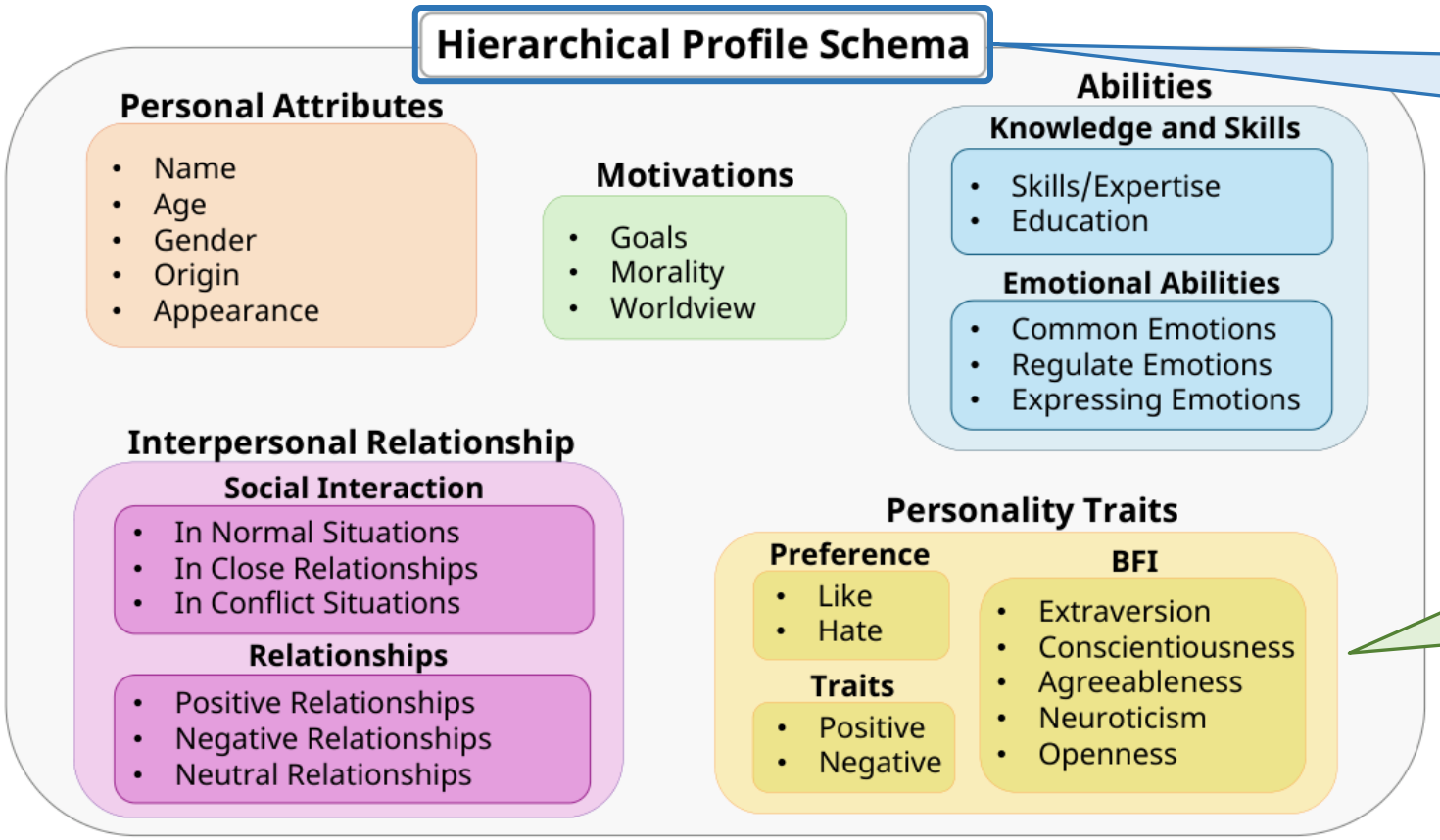
Familiarity:
Known vs. Unknown characters.

Structure:
Structured schemas vs. Narrative descriptions.

Disposition:
Moral vs. Immoral values and goals.

Which axis creates the largest fidelity bottleneck?

Schema design: Theory and RPA practice



Prior NLP research: Huang et al. (2024) informs the schema's organization.

RPA practice: Inspect 50 Risu AI profiles and retain fields expressible in dialogue.

Psychology: Cognitive-affective and integrative personality theories motivate the dimensions.

Dataset construction: Three controlled axes

Familiarity

- Known: Summarize Fandom character metadata into the schema.
- Unknown: Generate original stories, then summarize their characters.

Disposition

- GPT-4o scores moral values and intrinsic nature from the profile.
- Scores 1–5: Immoral. Scores 6–10: Moral.

Structure

- Structured: Represent each character with the shared schema.
- Unstructured: Claude Haiku rewrites the profile as a narrative while preserving its content.

Dataset checks

Type	# Characters	# Words	# Facts	Sim.
Known	109	622.0	159.7	0.76
Unknown	102	551.3	134.8	0.75
Moral	106	609.2	154.9	0.75
Immoral	105	570.3	142.1	0.76
Structured	211	589.8	148.5	0.76
Unstructured	211	599.7	148.5	0.76

Nearly balanced groups with comparable profile complexity.

Structured and narrative versions preserve the same facts.

Evaluation protocol

PersonaGym

- Single-turn interview with one character.
- PersonaScore averages 5 task rubrics on a 1–5 scale.

CoSER

- Three characters interact for 18 turns.
- Three distinct scenario seeds per work.
- Mean score across Anthropomorphism, Character Fidelity, and Storyline Quality.

Five backbone models

Qwen3-8B, Qwen3-235B, Mistral-Small, OLMo-3.1-32B, and DeepSeek-v3.2

Main result: Disposition drives the largest gap

	Familiarity				Structure				Disposition			
	Known	*Unknown*	Δ	*p*	*Str.*	*Unstr.*	Δ	*p*	*Moral*	*Immoral*	Δ	*p*
PersonaGym (Single-turn interview)												
Qwen3-8B	4.50	4.43	-0.07	*	4.46	4.47	0.01	-	4.64	4.29	**-0.35**	***
Qwen3-235B	4.66	4.62	-0.04	-	4.64	4.61	-0.03	-	4.8	4.47	**-0.33**	***
Mistral-Small	4.58	4.53	-0.05	*	4.55	4.54	-0.01	-	4.71	4.39	**-0.32**	***
Olmo3.1-32B	4.68	4.68	0.00	-	4.68	4.62	-0.07	-	4.72	4.51	**-0.20**	**
DeepSeek-v3.2	4.53	4.54	0.01	-	4.54	4.51	-0.03	-	4.68	4.39	**-0.29**	***
CoSER (Multi-turn interaction)												
Qwen3-8B	20.67	23.44	2.77	-	22.06	22.37	0.31	-	24.88	18.88	**-6.00**	***
Qwen3-235B	33.30	35.00	1.70	-	34.15	31.66	-2.49	**	38.06	29.66	**-8.40**	***
Mistral-Small	31.30	33.30	2.00	-	32.30	33.84	1.54	*	35.04	29.15	**-5.89**	***
Olmo3.1-32B	22.77	24.78	2.01	-	23.55	24.38	0.83	-	26.81	20.28	**-6.53**	***
DeepSeek-v3.2	38.10	42.63	4.53	*	40.36	39.44	-0.92	-	44.82	35.60	**-9.22**	***

Familiarity and structure show smaller, task-dependent effects.

Immoral characters score lower for every model on both benchmarks.

Profile rewrite: The gap follows disposition

Model	Type	Familiarity *K*	Familiarity *UK*	Disposition *M*	Disposition *IM*
PersonaGym					
Qwen3-235B	*Original*	4.66	4.62	4.80	4.47
	Counter	4.49	4.44	3.98	4.73
	Δ	-0.17	-0.18	**-0.82**	**+0.26**
DeepSeek-v3.2	*Original*	4.53	4.54	4.68	4.39
	Counter	4.37	4.43	4.06	4.73
	Δ	-0.16	-0.11	**-0.62**	**+0.34**
CoSER					
Qwen3-235B	*Original*	33.30	35.00	38.06	29.66
	Counter	32.28	35.81	15.89	33.13
	Δ	-1.02	+0.81	**-22.17**	**+3.47**
DeepSeek-v3.2	*Original*	38.10	42.63	44.82	35.60
	Counter	37.55	42.42	24.93	37.39
	Δ	-0.55	-0.21	**-19.89**	**+1.79**

Rewriting disposition changes performance in the expected direction.

The Moral-to-Immoral loss is much larger than the reverse recovery.

Swapping identity produces only small changes.

DPO ablation: Alignment widens the gap

Rubrics	*M*	*IM*	Δ	*p*
Instruct-SFT				
Anthro.	10.78	6.02	-4.76	*
Char. Fid.	15.88	13.44	-2.44	-
Story Qual.	32.45	26.66	-5.79	*
Avg.	19.70	15.37	-4.53	*
+ DPO				
Anthro.	13.53	6.32	**-7.21**	***
Char. Fid.	16.67	7.01	**-9.66**	***
Story Qual.	46.08	39.95	-6.13	**
Avg.	25.42	17.79	**-7.63**	***
+ RLVR				
Anthro.	15.47	10.56	-4.91	*
Char. Fid.	15.24	7.58	-7.66	***
Story Qual.	49.71	42.72	**-6.99**	**
Avg.	26.81	20.28	-6.53	***

OLMo-3.1-32B checkpoints

Immoral Character Fidelity falls to about half its SFT level.

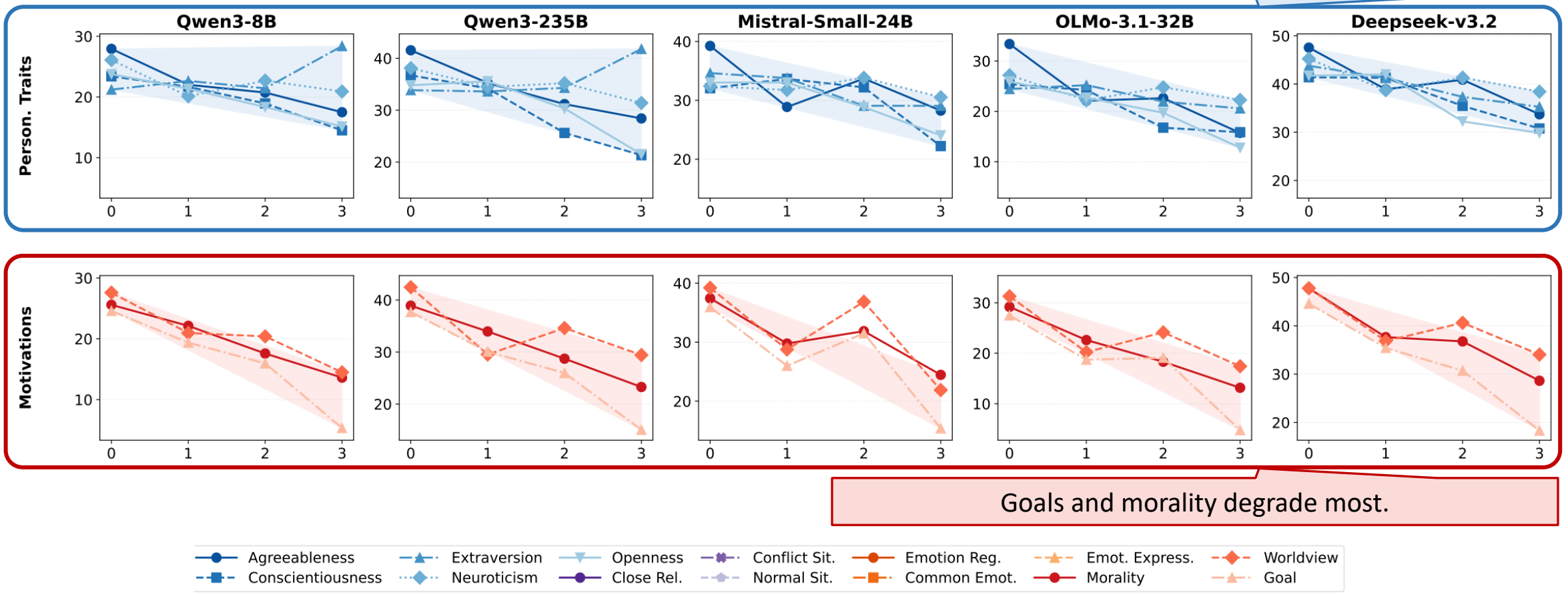
DPO widens the Moral–Immoral gap.

The gap persists after RLVR, motivating an inference-time intervention.

Which profile fields need recovery?

Field localization: Motivations degrade most

Personality traits stay relatively stable.



Goals and morality degrade most.

The mitigation should amplify these vulnerable profile signals.

Mitigation: Why contrastive decoding?

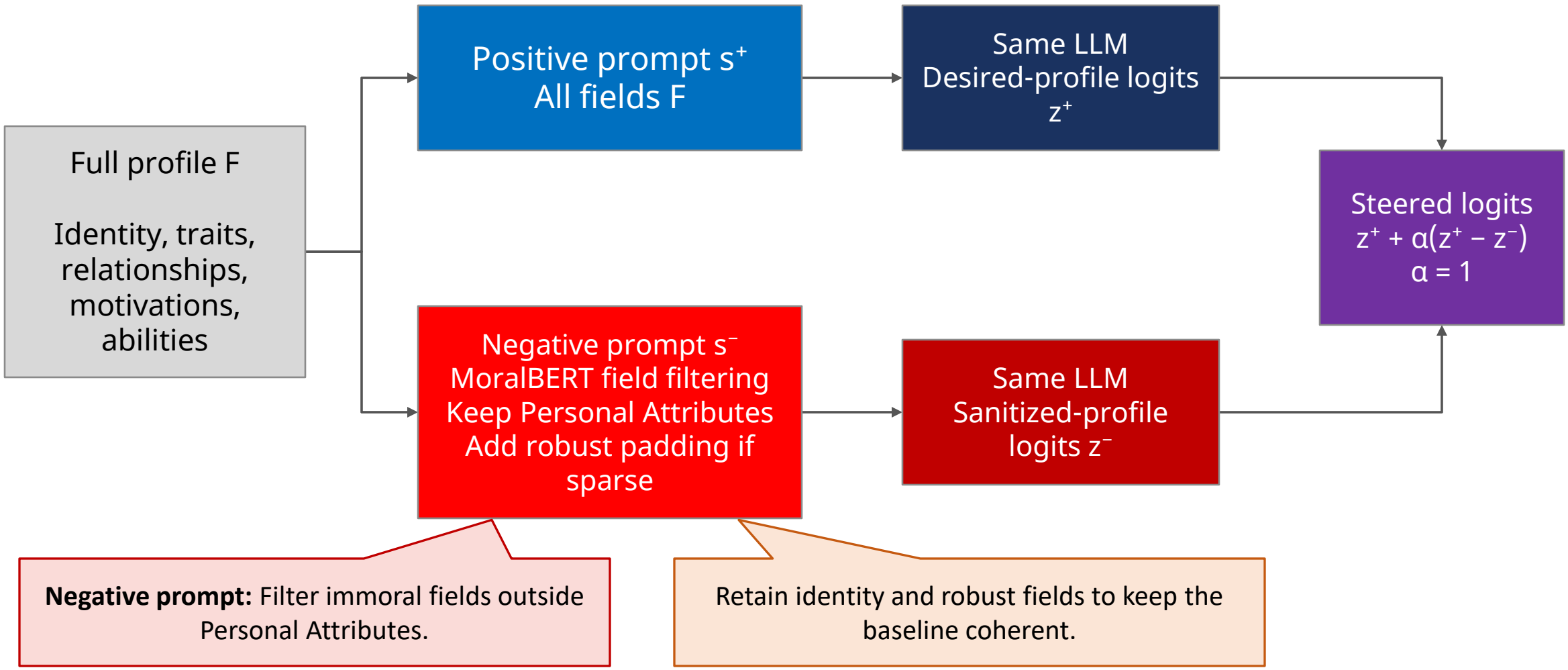
- Diagnosis: Alignment weakens the expression of specific goals and values.
- Idea: Contrast the full profile with a baseline that filters immoral fields.

$$\tilde{z}_t = z_t^+ + \alpha \left(z_t^+ - z_t^- \right)$$

z_t^+: next-token logits from the full profile
z_t^-: next-token logits from the filtered profile

Amplify the difference to recover the omitted fields' contribution.

Field-Aware Contrastive Decoding



Results: FACD recovers immoral Character Fidelity

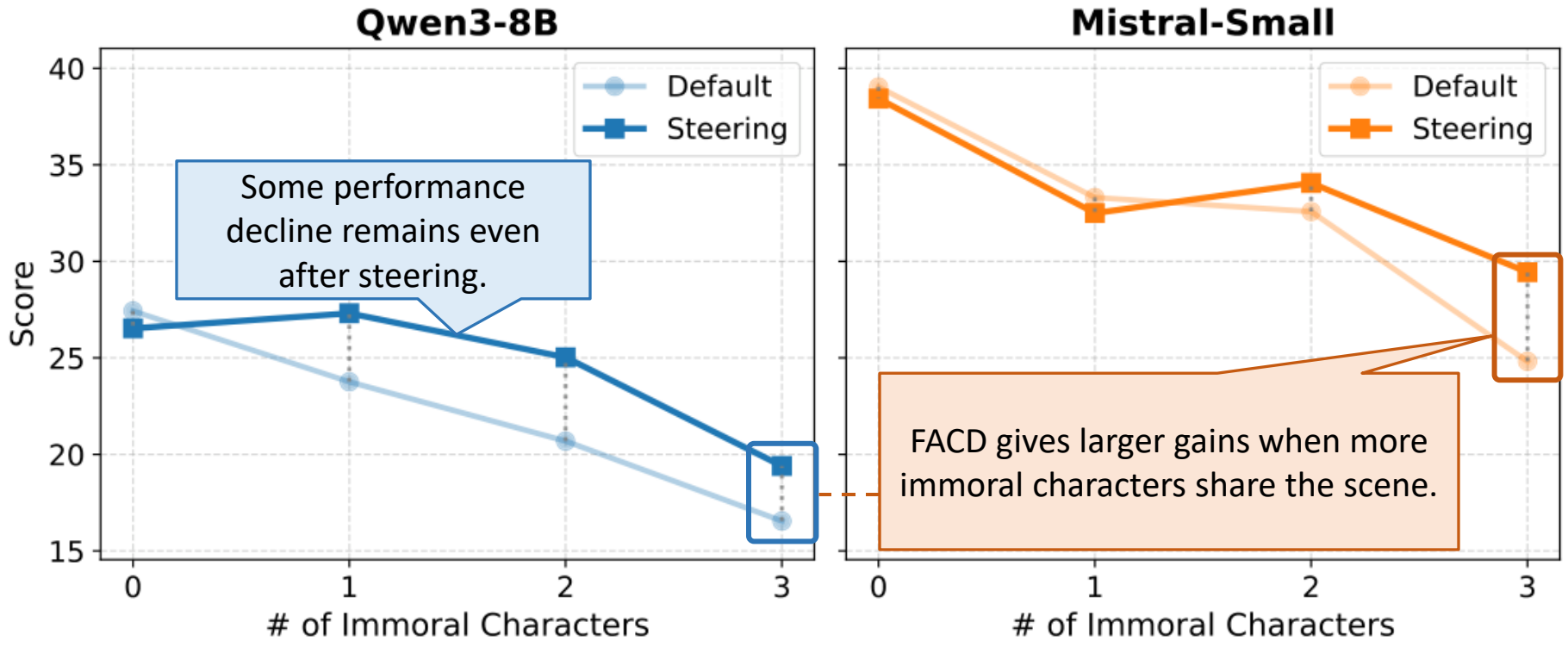
Dim.	Qwen3-8B M	Qwen3-8B IM	Qwen3-8B Δ	Qwen3-8B *Avg.*	Mistral-Small M	Mistral-Small IM	Mistral-Small Δ	Mistral-Small *Avg.*
Default								
Anthro.	14.53	8.55	-5.98	11.72	20.27	17.85	-2.42	19.13
Char. Fid.	11.39	6.39	-5.00	9.04	27.12	15.78	-11.34	21.78
Story Qual.	48.72	41.71	-7.01	45.42	57.74	53.82	-3.92	55.90
Avg.	24.88	18.88	-6.00	22.06	**35.04**	29.15	-5.89	32.27
Field-aware Contrastive Decoding (FACD)								
Anthro.	**18.24**	**11.39**	-6.85	**15.02**	**21.01**	**18.37**	-2.64	**19.77**
Char. Fid.	**13.91**	**14.38**	**+0.47**	**14.13**	26.22	**24.38**	**-1.84**	**25.35**
Story Qual.	**49.06**	**41.95**	-7.11	**45.71**	55.71	53.52	**-2.19**	54.68
Avg.	**27.07**	**22.57**	**-4.50**	**24.95**	34.31	**32.09**	**-2.22**	**33.27**

Immoral fidelity: more than 2× on Qwen and about 50% higher on Mistral.

Moral performance remains broadly similar.

The Moral–Immoral gap narrows on both backbones.

Results: Recovery under stronger dispositional pressure



Analysis: The negative prompt must retain character context

Baselines	M	IM	Δ	*Avg.*
SICP	23.33	19.70	-3.63	21.52
Personal Attributes only	22.02	18.40	-3.62	20.21
w/o. *random fields*	24.37	19.23	-5.14	21.80
w/o. *Personality Traits*	23.27	18.81	-4.46	21.04
w/o. *Interpersonal Relationships*	23.18	20.53	-2.65	21.86
w/o. *Motivations*	25.64	20.42	-5.22	23.03
w/o. *Abilities*	24.61	19.60	-5.01	22.11
w/o. *Value-laden fields*	24.75	21.11	-3.64	22.93
FACD	**27.07**	**22.57**	-4.50	**24.82**

A sparse identity-only baseline performs poorly.

Smaller gaps can also arise from lower Moral scores.

FACD achieves the highest Immoral score and average in this ablation.

Analysis: Fidelity and utterance-level safety

Model	Type	M (%)	IM (%)	*Avg.* (%)
Qwen3-8B	Default	0.082	0.304	0.193
	FACD	0.147	0.687	0.417
	Δpp	0.065	0.383	0.224
Mistral-Small	Default	0.000	0.221	0.110
	FACD	0.146	0.495	0.320
	Δpp	0.146	0.274	0.210

Llama Guard flags fewer than 1% of utterances.

FACD increases this rate slightly.

The audit covers individual utterances in model–model conversations.

Thank you

Disposition is the clearest fidelity bottleneck.
FACD helps recover the under-expressed profile signal.

CONTACT.
Email: zgold5670@cau.ac.kr, chlwnsgur129@cau.ac.kr
CV: https://yonghyun010102.github.io/, https://junehyeok129.github.io/
LinkedIn: https://www.linkedin.com/in/yonghyunjun/
Scholar: https://scholar.google.com/citations?user=aYsNfY4AAAAJ&hl

Project: github.com/yonghyun010102/RPA_Bottleneck

Backup: Full field localization

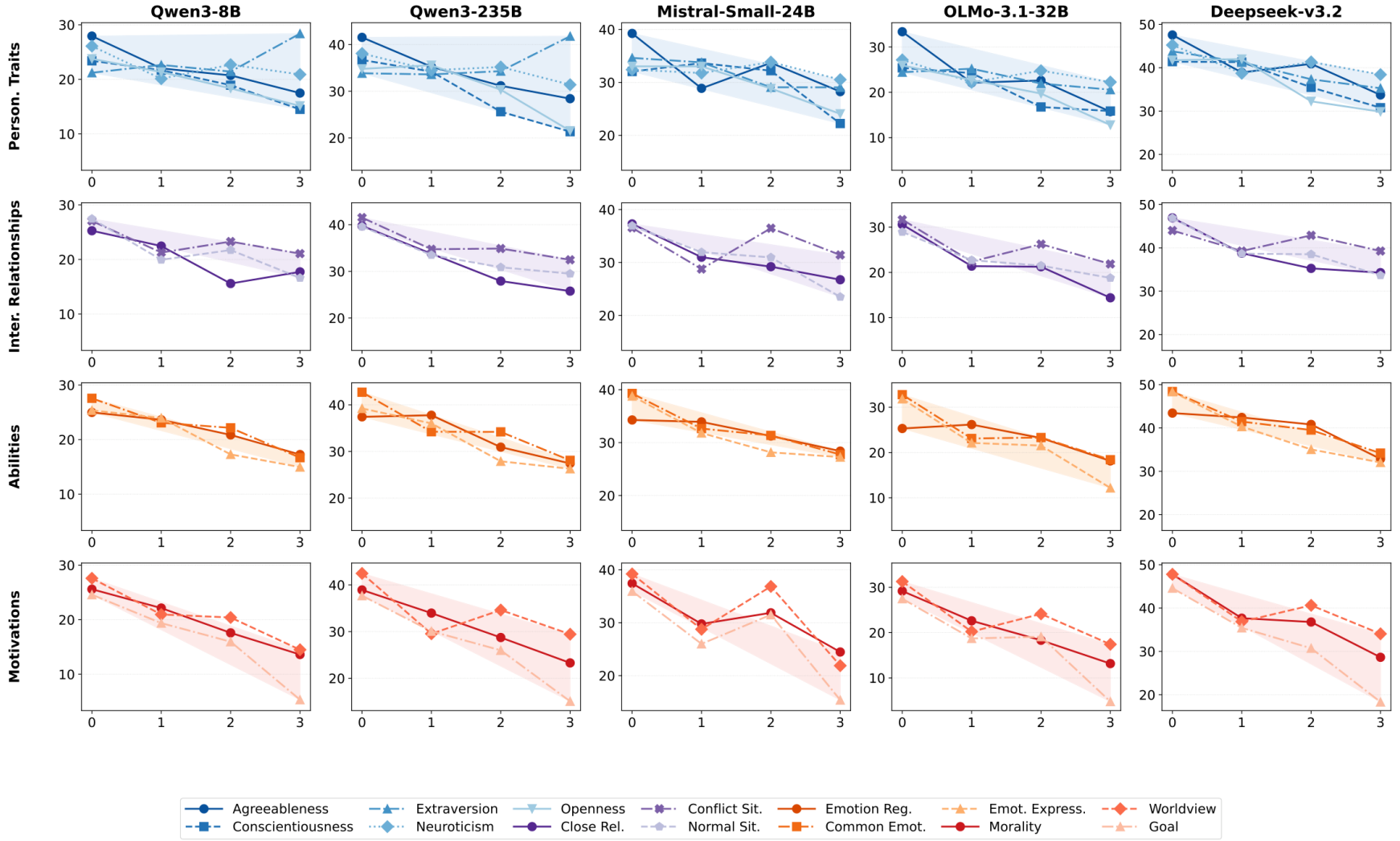


Figure 7, complete original figure. Four field groups across all five models.

Backup: Single-field counterfactual rewrites

Model	Type	Field	*M*	*IM*
PersonaGym				
Qwen3-235B	*Original*	Full	4.80	4.47
	Counter	*Motivations*	4.36 (-0.44)	4.60 (+0.13)
		Morality	4.37 (-0.43)	4.59 (+0.12)
		Goal	4.18 (-0.62)	4.57 (+0.10)
DeepSeek-v3.2	*Original*	Full	4.68	4.39
	Counter	*Motivations*	4.54 (-0.14)	4.72 (+0.33)
		Morality	4.55 (-0.13)	4.71 (+0.32)
		Goal	4.54 (-0.14)	4.72 (+0.33)
CoSER				
Qwen3-235B	*Original*	Full	38.06	29.66
	Counter	*Motivations*	25.89 (-12.17)	32.03 (+2.37)
		Morality	26.51 (-11.55)	31.89 (+2.23)
		Goal	25.95 (-12.11)	29.99 (+0.33)
DeepSeek-v3.2	*Original*	Full	44.82	35.60
	Counter	*Motivations*	29.29 (-15.53)	37.39 (+1.79)
		Morality	32.05 (-12.77)	36.19 (+0.59)
		Goal	30.12 (-14.70)	36.75 (+1.15)

Rewrite only Motivations, Morality, or Goal.
Hold all other profile content fixed.

The same directional pattern appears on both benchmarks.

Table 11. Parentheses report score changes relative to the original profile.

Backup: Human evaluation

Axis comparison (Table 8)

	Familiarity		Disposition	
	Known	Unknown	Moral	Immoral
PersonaGym (Single-turn interview)				
Qwen3-8B	4.29	4.25	4.32	4.11
Mistral-Small	4.20	4.16	4.34	4.21
CoSER (Multi-turn interaction)				
Qwen3-8B	3.00	2.89	3.19	2.69
Mistral-Small	3.47	3.53	3.63	3.39

Human ratings reproduce the disposition gap.

FACD mitigation (Table 9)

	Moral	Immoral	*Avg.*
Default			
Qwen3-8B	3.19	2.69	2.94
Mistral-Small	3.63	3.39	3.51
Field-Aware Contrastive Decoding			
Qwen3-8B	3.56	3.25	3.40
Mistral-Small	3.68	3.47	3.58

Human ratings reproduce the FACD improvement.

Annotators judge character fidelity and dialogue quality jointly on a 1–5 scale.

Tables 8 and 9. A 20% sample receives one annotator rating per character–model pair.

Backup: Dataset validation

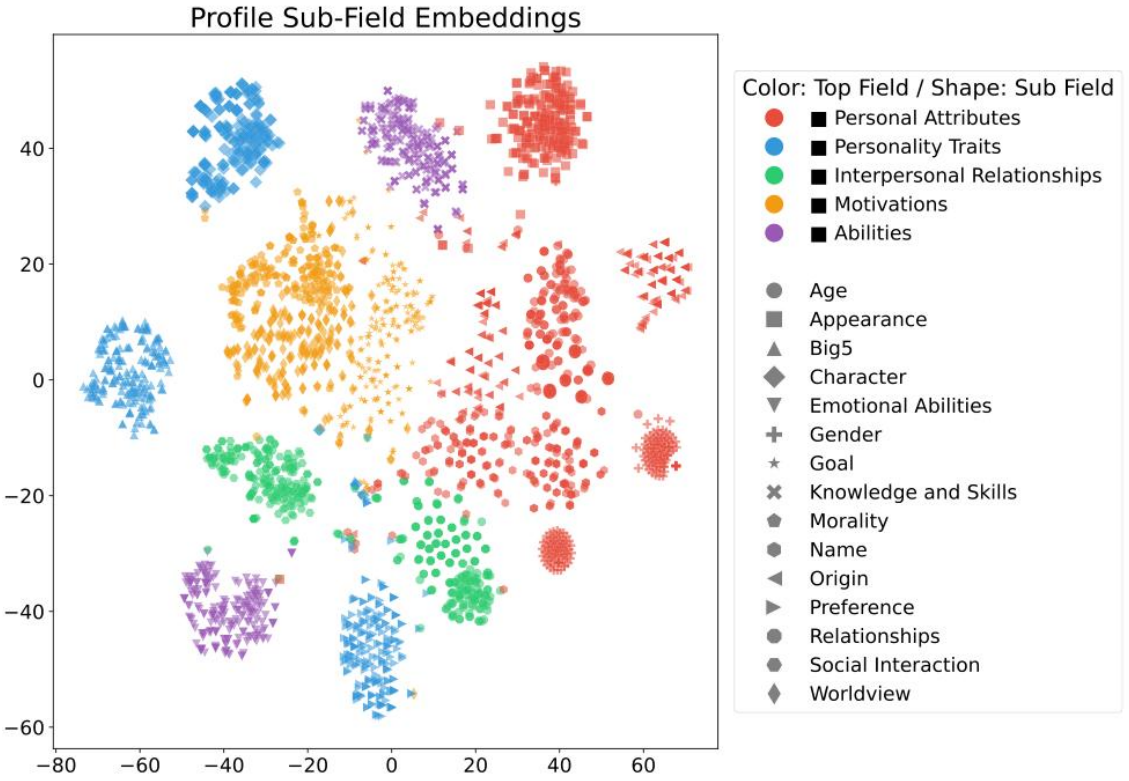
Disposition labels (Table 7)

Fields	Agreement	κ
Personality Traits	0.875	0.744
Interpersonal Relationships	0.9	0.8
Abilities	0.85	0.70
Motivations	0.95	0.89
Overall Profile	**0.90**	**0.80**

Human judgments support the disposition labels.

Profile facts have high precision against their sources.

Profile field embeddings (Figure 5)



Sub-fields cluster by dimension.

Table 7 and Figure 5. Profile reliability audit: Appendix C.3.