LET MINSTREL TELL YOUR STORY: RECONSTRUCTING CHARACTER PROFILES FROM DESCRIPTIONS FOR ROLE PLAYING

Anonymous authors
Paper under double-blind review

ABSTRACT

Role-playing agents increasingly stand in for a specific person or character, in digital twins, social simulation and expert decision support. The characters the field builds for are canonical figures from widely circulated fiction, and each arrives with a gold-standard profile. Users also want niche characters, original creations and people they know, none of which come with a profile or with an identity the backbone can draw on. Hiding the name of a known character does not close this gap, since removing the identity also removes most of the value the profile carried. We propose Minstrel, which rebuilds a profile from free description and delivers it identity-free. It poses three profiling questions for each of eight dimensions drawn from personality psychology, converts the material each question retrieves into personality content, and assembles one conditioning text. Across four open-source actors and two cross-family judges, every Minstrel configuration beats a no-information control. Built from public comments alone, the profile reaches win shares of 0.570 and 0.610 while the actor identifies the character in 1.6% of trials. It also keeps the sixteen characters as distinct as their official profiles do. The code are available at `https://anonymous.4open.science/r/Minstrel`.

1 INTRODUCTION

Role-playing agents are moving out of entertainment and into tasks built around one specific person or character. They stand in for an individual in a digital twin, populate social simulations (Park et al., 2023), and carry a domain expert's perspective into decision support. Each of these uses asks the system to sustain the personality of that one person. The characters the field actually builds for are narrower than that. They are canonical figures from literature, film and television, and they arrive with a gold-standard profile, abundant source material, and an identity the backbone may have met during pretraining. That availability has shaped system design and evaluation together. Prompting-based systems place a profile in context and rely on the backbone to act on it (Wang et al., 2024a; Li et al., 2023; Wang et al., 2025), and fine-tuned systems place the character inside the weights (Shao et al., 2023; Lu et al., 2024; Yu et al., 2024). The work that builds a profile rather than consuming one reads the canonical text itself (Yuan et al., 2024; Park et al., 2025). Daily demand is broader than this familiar catalog. A user may want a niche character, an original creation, or an ordinary person from their own life, such as a parent, a friend or a classmate. Such people can matter more to the user than a famous fictional figure, but they rarely come with a gold-standard profile,

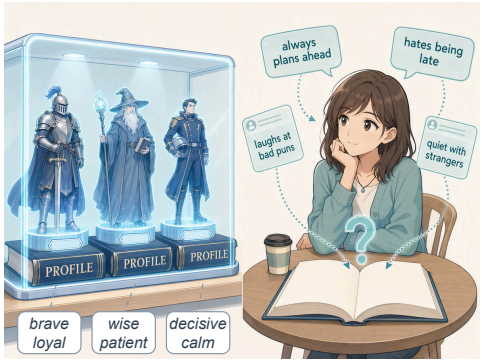

Figure 1: Who the field builds for and who users ask for. A canonical character comes with a written profile, whose headline attributes appear under each book, and an identity the backbone knows. The people users want come with scattered description and nothing else.

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.

Table 1: The task against its nearest neighbours.

Task	Input	Profile exists	Known to backbone	Profiled by
Role playing with a profile	written profile and name	yes	yes	nobody
Character profiling	canonical text	no	usually	system
Persona customisation	definition written for the purpose	by the user	varies	user
Unknown-character evaluation	written profile, name hidden	yes	yes	nobody
Profile reconstruction from description (ours)	description not written to be a profile	no	no	system

and Figure 1 shows the contrast. What exists for such a subject is description. People talk about how someone responds to pressure, treats others, makes decisions, and expresses emotion, and that material is plentiful but was never written to be a profile. We call the task profile reconstruction from description. Its input is a subject together with whatever free description of that subject exists, and its output is a conditioning text from which a generic actor can play the subject. Three conditions separate it from the tasks the field already has. There is no profile to retrieve or to redact. The description was not written to be a profile, so it is unstructured, redundant and partly contradictory. The profiling therefore has to be done by the system rather than by the user. The backbone has no entry for the subject, so the name alone pulls in nothing. The output therefore has to carry the whole character, and it has to be checked for doing so. Table 1 sets the task beside its nearest neighbours, none of which meets all three conditions. Persona customisation products let a user author a definition by hand, which moves the profiling onto the user and still consumes the result as a profile. Unknown-character evaluations hide the name of a character the backbone knows. Hiding the name is also the usual workaround, so we measured what it costs. We took the official profiles of sixteen canonical characters and removed identity in four steps, holding the content fixed. Removing identity costs 14.3 points of win share, which is most of the margin the profile carried. Actor-side recognition falls from 90.6% to 6.2% across the four levels, and the win share falls from 0.769 to 0.626 in step with it. Appendix A reports the full measurement. A subject the actor cannot place is therefore harder to play than the literature assumes.

That description is what a human profiler would work from, so we take it as the starting point and rebuild the profile from it. Minstrel, the method we propose, reaches a subject only through a fixed set of information needs, each one naming a single thing the profile has to establish. Every need is resolved against whatever description exists, and what comes back is converted into personality content rather than into events. The per-need outputs are then assembled under a fixed dimension scheme into one text the actor is conditioned on. The interface is the set of needs rather than any particular source. Retrieval from a corpus, elicitation from a language model and a user answering in person therefore become the same operation. Minstrel describes a subject along eight dimensions drawn from the structure personality psychology uses to describe a person. Each need is stated as one of 24 profiling questions, three per dimension. A question is one way to state a need, and the construction does not depend on it. Two interchangeable cue strategies are crossed with three description sources, a corpus of public comments, an official profile, and an external language model queried by name. The six configurations can then be compared on equal terms.

Across four open-source actors and two cross-family judges, all six configurations beat a no-information control. Reconstruction from public comments reaches win shares of 0.570 and 0.610 while the actor names the subject in 1.6% of trials. Personality ratings separate the sixteen subjects on every factor as sharply as the official profiles do. The separation reproduces those profiles dimension by dimension and in its pairwise distance structure. Against an official profile delivered at the same masking depth, Minstrel closes a substantial part of the distance. A dimension-level audit places the residue in contradiction absorbed from the source rather than in missing coverage.

Our contributions are as follows.

- We identify a problem class current role-playing research does not cover, in which the subject has no gold-standard profile and no identity the backbone can draw on. We also

show that stripping the identity from an existing profile removes most of the value the profile carried, so the usual workaround does not reach it.

- We propose Minstrel, which rebuilds a profile from free description rather than retrieving or redacting one. It fixes an interface of 24 profiling questions over eight personality dimensions and a chain of acquisition, conversion and assembly under five constraints. Its output is one identity-free conditioning text.
- We collect a comment corpus for sixteen characters and evaluate Minstrel against every requirement the problem sets, on four open-source actors and two cross-family judges. We report which design decisions carry the result and locate the residual distance to a gold-standard profile in disagreement absorbed from the source.

2 Related Work

Role-playing agents. Prompting-based systems place a character description in context and rely on the backbone to act on it. RoleGPT pairs a free-text profile with retrieved reference utterances (Wang et al., 2024a). ChatHaruhi separates a persona prompt from a vector store of character lines (Li et al., 2023). PersonaForge organises a character into core traits, speaking style and dynamic state, and generates through an inner-monologue stage (Tong & Zou, 2026). CoSER supplies structured character records together with an evaluation harness whose actor backbone is a runtime argument (Wang et al., 2025). All of them assume the profile already exists. Persona customisation products such as character.ai let a user author a definition by hand, which moves the profiling work onto the user and still consumes the result as a profile. A second family fine-tunes on character data, including Character-LLM (Shao et al., 2023), Ditto (Lu et al., 2024) and Neeko (Yu et al., 2024). Fine-tuned systems place the character inside the weights, which removes the possibility of supplying a new subject at inference time.

Building a profile rather than consuming one. Character profiling generates a profile from the source work and evaluates it against expert-written references (Yuan et al., 2024). CharacterGPT builds a persona incrementally from chapter summaries (Park et al., 2025). Both read the work itself, which presupposes a canonical text. Minstrel reads free description, which is what remains available for a subject with no canon. Psychological profiling supplies the structure we borrow, since practitioners work through a fixed question set rather than from an authoritative biography.

Evaluating character fidelity. CharacterEval scores single-turn Chinese dialogue continuations with a trained reward model (Tu et al., 2024). InCharacter administers psychological scales through interviews (Wang et al., 2024b). SocialBench targets multi-turn social interaction (Chen et al., 2024). RoleAgentBench supplies script-derived items and reports open-ended responses through pairwise comparison, which is the protocol we follow (Liu et al., 2024). Closest to our motivation is a controlled study that varies character familiarity together with profile structure and reports that familiarity has negligible effect (Jun et al., 2026). Our ladder in Appendix A indicates that this holds at name-level masking and changes once masking is deep enough to defeat actor-side recognition.

3 Minstrel

Minstrel builds a conditioning text for a subject from free description of that subject. We first define the problem and what a solution has to achieve, then describe the four stages of the construction, which Figure 2 shows end to end.

3.1 Problem definition

An actor is an LLM π that answers an item x under a conditioning text c, producing a response $y \sim \pi(\cdot \mid c, x)$. Role-playing systems assume that a subject s arrives with a gold-standard profile $c_s^\star$ that someone wrote for it in advance. We study the complement of that assumption. The subject is drawn from a population $\mathcal{S}$ that includes niche characters, original creations and ordinary people, so no canonical text and no catalog entry can be assumed. The available evidence is a set $\mathcal{D}_s$ of free descriptions of the subject, which is unstructured, redundant and partly contradictory. The

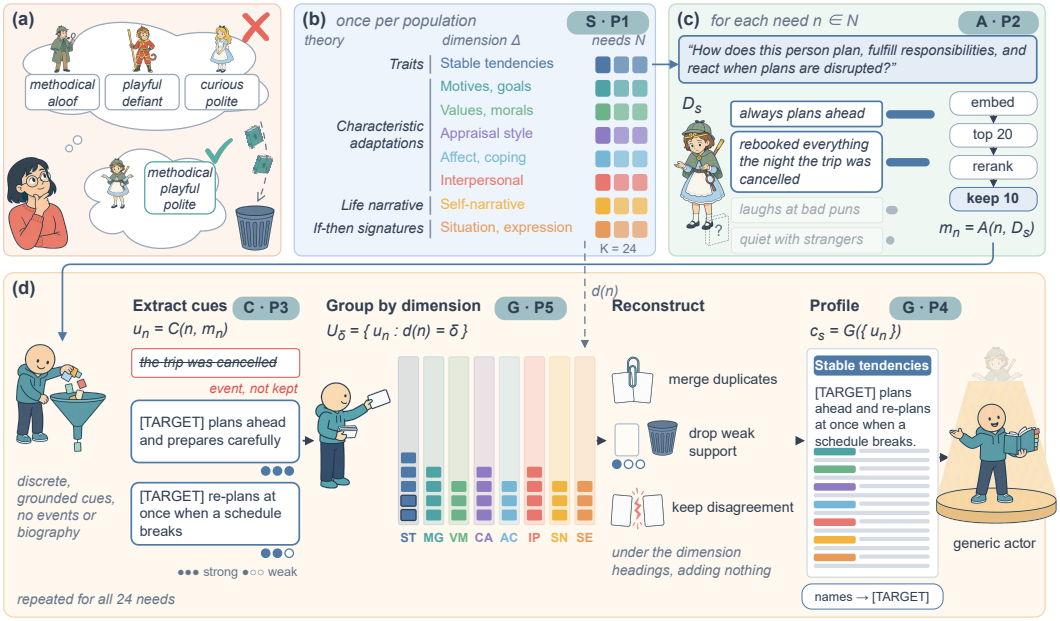

Figure 2: Overview of Minstrel and its four stages. (a) A user turns down familiar characters, each summarised by the headline attributes of its profile. The new character she settles on combines one attribute from each, and no existing profile describes it. (b) Specification, carried out once for a population. Levels of personality description (McCrae & Costa, 1997; McAdams & Pals, 2006; Mischel & Shoda, 1995) fix eight dimensions with three needs each, and each colour marks one dimension. (c) Acquisition for one need, which retrieves the relevant description by embedding search and reranking. (d) Cue conversion and profile reconstruction, shown with the extractor. The material becomes grounded cues with a support level, and the cues of all needs are grouped by dimension. Each group is reconciled and written into an identity-free profile that a generic actor plays. Stages carry their operators and principles. The subject and its description are illustrative.

backbone cannot be relied on to recognise s, so whatever reaches the actor has to carry the character by itself. Keeping the assumption and removing the identity from an existing profile, $c = \mathrm{mask}(c_s^\star)$, presupposes the object that is missing. Appendix A shows that it costs most of the profile's value even when the object exists. The task is therefore to construct c_s from $\mathcal{D}_s$ directly. What c_s has to achieve follows from the same properties. It has to help an actor that does not know the subject, and it has to keep different subjects apart rather than collapse them into one generic persona. Where a gold-standard profile exists, c_s should agree with it dimension by dimension. It has to carry nothing that identifies the subject, and the same construction has to work from whichever description exists. Sections 4.2 to 4.5 turn these requirements into measurements, and Appendix B states them as one constrained objective.

3.2 Specification of Information Needs

Minstrel represents a profile as the resolution of a fixed set of information needs $N = \{n_1, \ldots, n_K\}$. Each need names one thing the profile has to establish, and a map $d : N \to \Delta$ assigns it to one of M personality dimensions. Specification fixes both once for a whole population of subjects.

$$\mathcal{S} : \Delta \mapsto (N, d) \tag{1}$$

Neither N nor d takes the subject as an argument, which is what allows two subjects to be built the same way (P1). The dimension scheme follows the levels at which personality psychology describes a person, as Figure 2(b) shows. Stable tendencies correspond to traits (McCrae & Costa, 1997). Motives and goals, values and moral priorities, cognitive and appraisal style, affect and coping, and interpersonal patterns are characteristic adaptations (McAdams & Pals, 2006). They are the goals, beliefs, affects and self-regulatory plans that the cognitive-affective account of personality places between trait and behaviour (Mischel & Shoda, 1995). Self and narrative identity is the level of the

life narrative, and situation and expression captures the if-then signatures through which a stable personality shows differently across situations. Each need is posed as a natural-language profiling question, the form a human profiler already works with. One example asks how this person plans, fulfills responsibilities, and reacts when plans are disrupted. The questions ask about tendencies and the conditions under which they hold, not about events. Under situation and expression, one question asks which situations are exceptions to the person's general tendencies, which is the if-then signature stated as a question. There are three questions per dimension and $K = 24$ in all, so no dimension is carried by more evidence than another. Table 8 lists the full set. A slot in a schema or a retrieval query would serve the same role, and a user who wants a character that exists nowhere can answer the questions directly.

3.3 Acquisition

Acquisition $\mathcal{A}$ returns the material in $\mathcal{D}_s$ that bears on one need.

$$m_n = \mathcal{A}(n, \mathcal{D}_s) \tag{2}$$

It is applied once per need, so the description is never read as a whole and never decides for itself what is worth keeping. $\mathcal{A}$ has the same signature for every source, so a source can be swapped without touching the later stages (P2). For a corpus of description, Minstrel embeds the question, retrieves twenty candidate passages and reranks them to the ten it keeps, as in Figure 2(c). For a written profile or an external language model the question is put to that source instead, and Section 4.2 compares the three. Either source is asked to report observable tendencies and the evidence for them, using only what it supports. When it does not address a question, the answer says so and is marked unsupported rather than filled in.

3.4 Cue conversion

Conversion $\mathcal{C}$ turns the material for one need into personality content.

$$u_n = \mathcal{C}(n, m_n) \tag{3}$$

The content has to state how the subject tends to act rather than what happened to the subject, so $\mathcal{C}$ is not a copy of m_n (P3). Minstrel offers two interchangeable strategies for $\mathcal{C}$. The personality-cue extractor emits discrete, falsifiable cues, each grounded in the material, and it discards plot facts and biography. It looks for the kinds of content the dimensions name, namely dispositions, motives, values, appraisal style, affect regulation, relational patterns and expressive style. Each cue is graded strong, moderate or weak by how much of the material bears it out, and material that supports nothing yields no cues at all. Direct summarisation writes the same material as connected prose in answer to the question. Figure 2(d) shows the extractor, and Section 4.2 compares the two.

3.5 Profile reconstruction

Assembly $\mathcal{G}$ merges the per-need content into the text the actor receives.

$$c_s = \mathcal{G}\big(\{u_n\}_{n \in N}\big), \qquad U_\delta = \{\, u_n : d(n) = \delta \,\} \tag{4}$$

It collects U_δ for every dimension δ and writes one section per dimension, so every dimension is addressed and two subjects can be compared on identical axes (P5). With the extractor, duplicate cues within a section are merged and cues the evidence does not support are dropped. Under both strategies conflicting content is kept side by side rather than smoothed, and the text says that the evidence conflicts. Nothing is added that the input does not contain, and the result is one markdown document with a heading per dimension. The subject in this problem class has no identity the backbone could use, so any identifying signal in c_s would be an artefact of construction (P4). Every prompt that produces text the actor may receive therefore carries an identity rule. It refers to the subject only as a placeholder, removes proper nouns and searchable events, and generalises settings to their least specific form. Composing the stages gives the construction in one line.

$$c_s = \mathcal{G}\Big(\big\{\, \mathcal{C}\big(n, \mathcal{A}(n, \mathcal{D}_s)\big) \,\big\}_{n \in N}\Big) \tag{5}$$

Building one subject costs K applications of $\mathcal{A}$, K of $\mathcal{C}$ and one of $\mathcal{G}$. Appendix C gives every prompt, and Appendix B collects the five constraints P1 to P5 with the requirement each one serves.

4 EXPERIMENTS

4.1 SETUP

We evaluate on the open-ended response portion of RoleAgentBench (Liu et al., 2024). An item is retained under three criteria. Canonical knowledge absent from the actor's visible input must not be required, stable personality must bear on the answer, and the visible input must determine the answer. Applying all three leaves 95 items over ten characters. The comment corpus holds 8,673 public comments about 16 subjects, written by 6,569 distinct authors across six platforms. English accounts for 5,104 of them and Chinese for 3,569. Two factors define a configuration. The cue strategy is the extractor or direct summarisation of Section 3.4, and the source is the comment corpus, the official profile, or an external language model queried by name. Crossing them gives six configurations, which is how P2 is tested rather than assumed. Routing the official profile through the identical pipeline separates the effect of the source from the effect of the pipeline. Canonical characters are used so that an official profile exists as a reference, which means acquisition sees a name where a deployment would not. The conditioning text the actor receives still carries none of it, and Section 4.5 verifies that with an actor-side probe. Four open-source actors are used, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Appendix D gives the retrieval settings, the corpus composition and every model revision. Actors are served with vLLM (Kwon et al., 2023) on one GPU per job on a Slurm cluster (Yoo et al., 2003). The devices are H200, H100 NVL, L40S and A100, with 8 CPU cores and 96 GB of host memory per job.

Conditioning texts are built by gpt-5.6-sol. Responses are compared pairwise against the benchmark reference by two judges from other families. Judge A, grok-4.3, decides every comparison, and Judge B, Kimi-K2.6 (Kimi Team, 2025), repeats those against the control and the de-identified profile and between cue strategies. Every pair is judged in both display orders, since LLM judges favour one position (Zheng et al., 2023). A pair whose verdict changes with order counts as undecided, and judges may also declare a tie. We report the winner's share of decided pairs with a Wilson interval (Wilson, 1927), and use an exact sign test with Holm correction (Holm, 1979) within each family of contrasts. The judging set comprises 113,971 verdicts. Before any condition is compared, a de-identified official profile must beat the no-information control. Both judges satisfy this, with win shares of 0.626 and 0.686 and intervals [0.593, 0.658] and [0.655, 0.715].

4.2 UTILITY AND PORTABILITY

Utility is the share of decided pairs in which a judge prefers the response under c over the response under a no-information control $\emptyset$ on the same item.

$$W(c) = \Pr_{\pi\sim\Pi,\, x\sim X} \big[\, y_c \succ y_\emptyset \,\big], \qquad y_c \sim \pi(\cdot \mid c, x) \tag{6}$$

Here Π is the set of actors, X the set of items and $\succ$ a judge verdict. The comparator is a no-information control rather than another system, so the quantity measured is the value of the conditioning itself. A configuration helps when the interval on W excludes $1/2$, and portability requires that under every source. Table 2 reports every configuration against the no-information control. All twelve intervals, six configurations under two judges, exclude 0.5. Reconstruction from public comments reaches 0.570 and 0.610 while the actor recognises the subject in 1.6% of trials. Both judges order the sources identically, official profile above external language model above comments, so P2 holds for all three sources the instance claims. Routing the official profile through Minstrel leaves it statistically equivalent to the same profile delivered directly, with intervals [0.455, 0.526] and [0.487, 0.558]. The pipeline neither adds nor destroys usable content, so the distance the comment configurations still carry is attributable to the source rather than to the processing. Figure 3(a) breaks the comment configurations down by actor. Mistral-7B-Instruct-v0.3 gains most and Qwen2.5-7B-Instruct gains least, with an interval that includes parity under both cue strategies. A reconstructed profile is worth more to an actor with less of its own material to fall back on. Holding the source fixed, the extractor does not improve on direct summarisation. Win shares for the extractor are 0.462, 0.446 and 0.476 under Judge A across the three sources, and 0.434, 0.454 and 0.463 under Judge B. The non-decision rate for this family is the highest in the study, between 27% and 32% against 20% to 25% elsewhere. The extractor retains an advantage in identity leakage, yielding lower actor recognition across all three sources, and it contradicts the official profile more often, 10.2% against

Table 2: Each Minstrel configuration against the no-information control.

		Judge A		Judge B	
Cues	Source	Share	95% CI	Share	95% CI
Extractor	Comments	0.570	[0.535, 0.604]	0.610	[0.577, 0.642]
Summary	Comments	0.563	[0.529, 0.597]	0.622	[0.589, 0.654]
Extractor	Profile	0.629	[0.595, 0.662]	0.635	[0.603, 0.667]
Summary	Profile	0.644	[0.610, 0.676]	0.654	[0.622, 0.685]
Extractor	LLM	0.608	[0.574, 0.641]	0.630	[0.598, 0.661]
Summary	LLM	0.620	[0.587, 0.652]	0.649	[0.617, 0.680]

Table 3: Between-subject separation of personality ratings, intraclass correlation.

Rated material	O	C	E	A	N
Official profile	0.925	0.882	0.957	0.961	0.824
Minstrel profile, Extractor	0.866	0.952	0.966	0.940	0.884
Minstrel profile, Summary	0.860	0.877	0.887	0.910	0.905
Actor behaviour, Extractor	0.844	0.769	0.882	0.749	0.764
Actor behaviour, Summary	0.914	0.758	0.888	0.862	0.571

4.7%. Direct summarisation is the better default, and the extractor is preferable when leakage matters more than fidelity. The factor that does carry the result is the source, and its ordering is stable across judges. Both cue strategies therefore satisfy P3, and the open problem sits upstream of them in the source.

4.3 Distinctiveness

Separation is measured by rating every subject on the five factors of the five-factor model (McCrae & Costa, 1997) at three points in the pipeline. These are the official profile, the Minstrel reconstruction, and the actor responses produced under that reconstruction. Each cell is rated three times by a model that built none of the material, giving a within-cell standard deviation of 0.63 on a nine-point scale. A one-way analysis of variance over subjects then asks whether differences between subjects exceed the noise of repeated rating. For factor f we report the intraclass correlation (Shrout & Fleiss, 1979), the share of rating variance that lies between subjects.

$$\rho_f = \frac{\sigma^2_{b,f}}{\sigma^2_{b,f} + \sigma^2_{w,f}} \tag{7}$$

The within-cell term $\sigma^2_{w,f}$ comes from the repeated ratings, and the reference is the value the official profiles attain on the same factors. Table 3 reports the result. Every cell is significant with p below 1.4×10^{-5}. The intraclass correlations for the Minstrel profiles run from 0.86 to 0.97, against 0.82 to 0.96 for the official profiles. Reconstruction from public commentary separates subjects as sharply as the gold-standard profiles do. A second probe asks the same question behaviourally. A judge is shown one response together with two anonymised profiles, the speaker's own and a distractor from a different subject, and picks the profile the speaker matches. Both profile orders are shown and order-dependent verdicts are discarded. Matching reaches 0.965 with interval [0.940, 0.980] for the extractor and 0.958 with interval [0.931, 0.974] for direct summarisation.

4.4 Fidelity

We test the separation against the official profiles and against independent type labels. Figure 3(b) reports the comparison dimension by dimension. The reconstructed profiles correlate with the official profiles at r of 0.63 on conscientiousness, 0.77 on extraversion and 0.57 on agreeableness, all significant across the sixteen subjects. Openness and neuroticism do not reach significance at the profile stage, with r of 0.17 and 0.31. In the actors' behaviour all five factors align, with r from 0.58 to 0.88 and a mean of 0.73. Correlating the 45 between-subject distances in behaviour against the

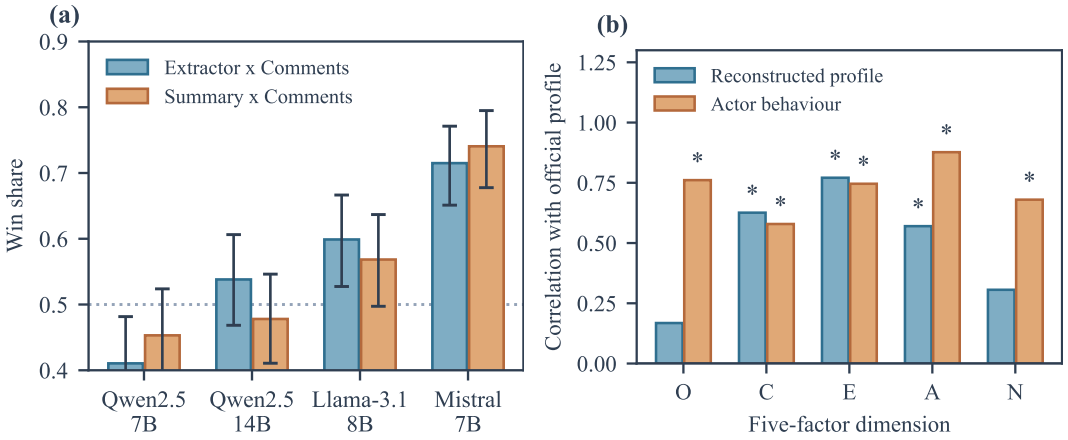

Figure 3: (a) Comment-sourced reconstruction against the no-information control by actor, under Judge A, with Wilson intervals and the dotted line at parity. (b) Per-factor correlation with the official profiles for the extractor configuration sourced from comments, with stars marking p below 0.05 across the sixteen subjects.

Table 4: Dimension-level agreement with the official profile.

Cues	Source	Full	Partial	None	Contradiction
Extractor	Comments	0.234	0.688	0.000	0.102
Summary	Comments	0.273	0.633	0.000	0.047
Extractor	Profile	0.688	0.266	0.008	0.000
Summary	Profile	0.758	0.211	0.000	0.000
Extractor	LLM	0.500	0.453	0.000	0.008
Summary	LLM	0.483	0.425	0.000	0.000

official profiles gives r of 0.541 with p of 2.5×10^{-5}. The geometry of who resembles whom is preserved, not only the marginal scores. Behaviour aligns more closely than the profile it was conditioned on, 0.73 against 0.49 on the dimension means and 0.541 against 0.162 on distance structure. Cues that read as partial in a written profile appear to be resolved into consistent conduct once an actor has to speak from them. We report the pattern as an observation, since our design does not isolate its mechanism. Independent MBTI labels (Myers et al., 1985) from Personality Database (Personality Database, 2026), available for 13 of the 16 subjects, separate the extraversion groups at every stage. Agreeableness separates in the official profile and in behaviour, while the conscientiousness axis weakens along the pipeline, from 0.022 in the official profile to 0.457 in behaviour. Appendix E reports the Welch tests, an Enneagram analysis and a lower-confidence comparison with Big Five Traits votes.

Against the official profile delivered at the same masking depth, the comment configurations sit below it with intervals [0.381, 0.449] and [0.415, 0.485]. We locate that distance with a dimension-level audit, which compares each reconstructed conditioning against the official profile on the eight personality dimensions. The auditing model built none of the material and was instructed to ignore proper nouns. It scores each dimension as full, partial, absent or contradictory, which gives coverage and contradiction over the dimension set Δ.

$$\mathrm{cov}(c) = 1 - \frac{|\{\delta \in \Delta : \text{no content}\}|}{|\Delta|}, \qquad \mathrm{con}(c) = \frac{|\{\delta \in \Delta : \text{conflicting content}\}|}{|\Delta|} \tag{8}$$

Coverage of every dimension is required. Contradiction is not required to be zero, but it has to be located, because where it comes from decides what to fix next. Table 4 reports both per configuration. Coverage is complete across every configuration, so comment-sourced reconstructions address every dimension and differ in precision rather than in scope. Contradiction separates the sources sharply. Comment-sourced reconstructions disagree with the official profile on 4.7% to 10.2% of dimensions, while the other four configurations stay at or below 0.8%. Public commentary contains

Table 5: Comparison against released systems, with the information level of each row.

Our condition	Information level	Share	95% CI
Identity-present profile	Profile and identity	0.645	[0.614, 0.676]
Masked official profile	Profile only	0.471	[0.438, 0.504]
Extractor x Comments, strongest backbone	Reconstruction only	0.543	[0.477, 0.607]
Extractor x Comments, all backbones	Reconstruction only	0.396	[0.364, 0.429]
Summary x Comments, all backbones	Reconstruction only	0.420	[0.388, 0.453]

disagreement about a subject and the pipeline absorbs it. The material itself is there, since 59.4% of the 3,740 retrieved comments describe what the person is like, and Appendix E reports the density analysis in full. The remaining distance is a matter of resolving disagreement in the source, which points at consensus weighting and conflict detection as the next step.

4.5 Concealment

P4 requires that the conditioning the actor receives carries no identity, so we ask the actor itself to name the character and the work, with an explicit unknown answer allowed.

$$\varepsilon(c) = \Pr_{\pi \sim \Pi} \left[R_\pi(c) = \mathrm{id}(s) \right] \tag{9}$$

Here R_π is the adversarial probe of Appendix C run on the actor that consumes the text. The reference $\varepsilon_0 = 0.062$ is the recognition a deeply masked official profile leaves behind. Actor recognition across the six configurations ranges from 1.6% to 7.8%, comparable to the deeply masked official profile at 6.2%. The extractor yields consistently lower recognition than direct summarisation, between 1.6% and 4.7% against 3.1% and 7.8%. A profile rebuilt from public commentary therefore reaches the actor without telling it who the subject is. We recommend reporting this probe alongside any unknown-character condition, for the reason Appendix A gives.

4.6 Position against released systems

Cross-method comparison is meaningful when the methods receive comparable information, so Table 5 states the information level of every row against two released training-free systems on the same items. PersonaForge receives the official profile together with the identity, which places it two levels above the Minstrel grid, and it runs on the same four backbones. At the matched level the same official profile delivered directly reaches 0.645 with interval [0.614, 0.676] against PersonaForge, so it outperforms the three-layer persona pipeline that consumes it. One level down, the masked official profile is already equivalent to PersonaForge with interval [0.438, 0.504], even though PersonaForge still holds the identity. Two levels down, on the strongest backbone, comment-only reconstruction is equivalent to PersonaForge with a win share of 0.543 and interval [0.477, 0.607]. A profile rebuilt from public commentary, delivered to an actor that recognises the subject in 1.6% of trials, matches a system holding the official profile and the name. Against CoSER our conditions lead with win shares of 0.827 and 0.852, with intervals [0.802, 0.850] and [0.828, 0.873]. We report that row as a reference point, since the backbone is unmatched and the checkpoint was trained on a different distribution.

5 Conclusion

Everyday role playing keeps asking for subjects that were never profiled. The field's workaround of hiding a known character's name does not reach that case, since the identity it hides was carrying most of the profile's value. We have treated the missing profile as something to be constructed rather than retrieved. Minstrel fixes a set of information needs, resolves each against whatever description exists, and delivers the result identity-free. It meets every requirement the problem sets, on four actors and two judges. Nothing in the four-stage construction depends on the particular question set, the retrieval stack, or the two cue strategies we compared. An instance that replaced all three would be judged on the same five criteria. What limits the instance today is disagreement in the source rather than anything the pipeline fails to cover, so consensus weighting and conflict detection are the next thing to build.

AI USE STATEMENT

Large language models are components of the artifact and of its evaluation rather than writing aids. One hosted model answers the profiling questions and assembles every conditioning text. Two further hosted models from different families serve as judges. Judge A also runs the separability probe, and Judge B carries the personality ratings and the dimension-level audit. Four open-source models act as the role-playing actors. Appendix D states which model plays which role and pins the open-source revisions, and Appendix C gives every prompt verbatim. We take responsibility for the final content of this work.

ETHICS STATEMENT

The description source is public commentary about fictional characters. Author identifiers are hashed at collection time and no comment is reproduced with an account name attached. No human subjects were recruited and no private communication was collected. The method itself carries a dual-use concern that we state plainly. A pipeline that rebuilds a usable personality profile from free description will work on a real person as readily as on a fictional one. The identity-free delivery that protects evaluation validity also removes the most obvious signal that a real person is being imitated. We evaluate only on fictional subjects, and we would expect any deployment on real people to require consent from the person being modelled.

REPRODUCIBILITY STATEMENT

Appendix C reproduces every prompt that drives the pipeline, including the identity rule appended to all output-facing stages. Appendix D reports the composition of the comment corpus, the retrieval settings, the role of each model, the pinned revisions of the four open-source actors, and the compute used for generation. Appendix A reports the measurement behind the problem statement. Every number in the text, in the tables and in the figures is read from a single frozen fact table. The figure scripts read that table directly, so a figure cannot drift from the prose. The code, the comment corpus, every conditioning text, the judge verdicts and the fact table are available at `https://anonymous.4open.science/r/Minstrel`.

REFERENCES

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, 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*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Yonghyun Jun, Junhyuk Choi, Jeonghyun Park, Jihyeong Park, Liu Nicole Geumheon, and Hwanhee Lee. Identifying and mitigating bottlenecks in role-playing agents: A systematic study of disentangling character profile axes. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*, 2024.

Dan P. McAdams and Jennifer L. Pals. A new big five: Fundamental principles for an integrative science of personality. *American Psychologist*, 61(3):204–217, 2006.

Robert R. McCrae and Paul T. Costa. Personality trait structure as a human universal. *American Psychologist*, 52(5):509–516, 1997.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268, 1995.

Isabel Briggs Myers, Mary H. McCaulley, and Robert Most. *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, 1985.

OpenAI. New embedding models and API updates. `https://openai.com/index/new-embedding-models-and-api-updates/`, 2024.

Jeiyoon Park, Chanjun Park, and Heuiseok Lim. Charactergpt: A persona reconstruction framework for role-playing agents. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track*, 2025.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Personality Database. Personality database. `https://www.personality-database.com`, 2026. Community-voted personality type labels, accessed 21 September 2026.

Don Richard Riso and Russ Hudson. *The Wisdom of the Enneagram : the Complete Guide to Psychological and Spiritual Growth for the Nine Personality Types*. Bantam Books, 1999.

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*, 2023.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Jizhou Tong and Sirui Zou. Personaforge: Psychology-grounded dual-process architecture for personality-consistent role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7845–7874, 2026.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.638. URL `https://aclanthology.org/2024.acl-long.638/`.

Noah Wang, Z.Y. 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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the 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 *International Conference on Machine Learning*, 2025.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Andy B. Yoo, Morris A. Jette, and Mark Grondona. SLURM: Simple Linux utility for resource management. In *Job Scheduling Strategies for Parallel Processing*, volume 2862 of *Lecture Notes in Computer Science*, pp. 44–60, 2003.

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*, 2024.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.

A Problem validation: what hiding identity costs

The problem definition in Section 3.1 rests on the claim that redacting an existing profile is not a route to a subject without one. This appendix reports the measurement behind that claim.

We hold profile content fixed and vary how far identity is removed. Level L0 keeps the official profile with the character's real name. Level L1 replaces the character's own name and aliases with a placeholder. Level L2 additionally replaces every proper noun through a hand-written lexicon covering other characters, places, institutions and setting vocabulary. Level L3 rewrites the profile so that plot events, signature abilities, catchphrases and mappable configurations of supporting characters are generalised away. Dispositions, motives, values, appraisal style, affect, relational patterns and expressive style are preserved.

Masking depth is verified against the actor, since the actor is the model whose memory the masking is meant to cut off. We ask each actor to identify the character and the source work from the

masked profile alone, allowing an explicit unknown answer. Table 6 and Figure 4 report recognition alongside downstream win share over a no-information control.

Table 6: Masking depth against actor recognition and downstream benefit.

Level	Masking	Char. ID	Work ID	Win share	95% CI
L0	Name present	0.906	0.812	0.769	[0.740, 0.795]
L1	Own name replaced	0.641	0.719	0.748	[0.718, 0.776]
L2	All proper nouns replaced	0.375	0.641	0.697	[0.665, 0.727]
L3	Deep rewrite	0.062	0.172	0.626	[0.593, 0.658]

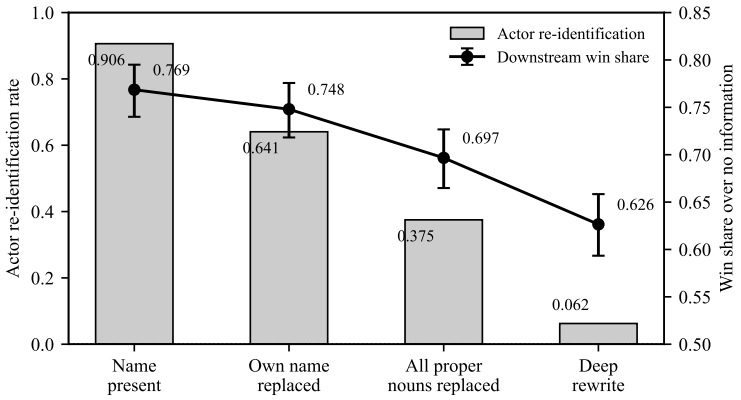

Figure 4: Masking depth against actor recognition and downstream win share.

Two readings follow. Hiding identity costs 14.3 points of win share once it is done thoroughly. That is most of the margin the profile carried in the first place. Redaction is therefore a poor route to a non-canonical subject. The second reading concerns how identity survives. Name replacement costs only 2.1 points and still leaves the actor naming the character two times out of three. Identity is carried by the configuration of a profile rather than by its names. For one character, two rounds of generalising individual phrases left recognition at 75%. The third round removed a set of five simultaneously mappable supporting roles. These were a strict senior teacher, a rough uneducated staff member, a platonic sibling-like friend, a bickering companion and a timid classmate. That round brought recognition to zero. Any one of those roles is generic. Five of them arranged around one subject form a cast list, and a cast list names a work.

Level L3 gives us more than a motivation. It gives us an experimental condition. An actor that recognises the subject in 6.2% of trials is, for the purposes of role playing, facing a character it does not know. The evaluation in Section 4 runs at that depth, which lets us test Minstrel under non-canonical conditions using subjects for which ground truth exists.

Identity as a second channel. The ladder removes identity from a profile. The same question can be asked from the other direction, by supplying identity without a profile. Supplying only the subject's name and work, with no profile, beats the no-information control with a win share of 0.734 and interval [0.704, 0.763]. That exceeds the deeply masked official profile at 0.626. Restoring identity to a fixed profile content raises the win share to 0.706 with interval [0.675, 0.736] over the masked version. Identity and profile content are therefore two channels rather than one, and for widely circulated characters the actor's own memory is a strong channel on its own. The subjects Minstrel is designed for lack this channel, which is why the evaluation runs at L3 rather than at the masking depth the field currently uses.

Consequence for evaluation practice. A study that constructs an unknown-character condition by hiding the name is working at the first masking level. At that level the actor still recognises the subject two times out of three, and the benefit of the profile is nearly intact. We recommend

reporting actor-side recognition alongside any unknown-character condition, so that readers can tell which masking depth a result describes.

B Design constraints and objectives

Table 7 collects the five constraints that Section 3 places on the stages, together with the requirement of Section 3.1 that each one serves. P1, P2 and P5 follow from the interface, since $\mathcal{A}$ has the same signature for every source and grouping by $d(n)$ lets coverage be checked one dimension at a time. P3 and P4 constrain an operator, conversion and assembly respectively.

Table 7: The five constraints on the stages of Minstrel, with the requirement each one serves.

Constraint	Stage	Design decision	Requirement
P1 Fixed elicitation interface	Specification	Represent the profile as the resolution of a fixed set of information needs rather than as a free document	Portability across subjects
P2 Source substitution	Acquisition	Resolve every need against whichever source exists, a corpus, a model, or the user	Utility under each source
P3 Personality first	Conversion	Convert material into dispositions and patterns rather than events and biography	Fidelity to a reference profile
P4 Identity-free delivery	Assembly	Deliver a text that carries nothing which identifies the subject	Actor-side recognition
P5 Dimensional assembly	Assembly	Assemble per-need output under a fixed dimension scheme	Coverage and distinctiveness

The requirements can be stated as one constrained problem over the operators, using the measures defined in Sections 4.2 to 4.5. The construction maximises utility subject to concealment and coverage.

$$\max_{\mathcal{A},\mathcal{C},\mathcal{G}} W(c_s) \qquad \text{subject to} \qquad \varepsilon(c_s) \le \varepsilon_0, \quad \mathrm{cov}(c_s) = 1 \tag{10}$$

A violation of either constraint disqualifies the construction, however large the utility is. The measurements are read in a fixed order, since each is only worth reading if the one before it held. Utility comes first, and distinctiveness next because utility alone is satisfied by a generic persona. Fidelity follows since distinct is not the same as right, and concealment closes the sequence because it can disqualify a construction that passed everything else. Ahead of all of them the evaluation surface itself is validated, by requiring that a gold-standard profile stripped of identity beats the no-information control.

C Prompts

The twenty-four profiling questions are the elicitation interface, and Table 8 gives them in full. Each is put to the description source once, in the English form below.

Every stage of the pipeline is driven by a fixed prompt. The identity rule below is appended to every prompt that produces text the actor may receive.

```
IDENTITY RULE (mandatory): your output must not allow a reader to identify
the person. Never write any proper noun -- no personal names, place names,
institutions, organisations, work titles, or franchise-specific vocabulary.
Never describe searchable plot events, signature possessions, unique
abilities, or catchphrases. Never describe a configuration of supporting
characters that could be mapped role-for-role onto a known cast. Refer to
the person only as [TARGET]. Generalise settings to their least specific
form. Psychological content is what matters: keep it, and strip everything
that anchors it to a specific work.
```

Table 8: The twenty-four profiling questions, three for each of the eight dimensions.

Dimension	Question
Stable tendencies	How does this person plan, fulfill responsibilities, and react when plans are disrupted? What stable differences appear with familiar people versus strangers? How does this person usually approach rules, uncertainty, novelty, and risk?
Motives and goals	What does this person repeatedly pursue or protect? What does this person most fear losing or try to avoid? For which goals will this person accept substantial personal cost?
Values and moral priorities	When loyalty conflicts with rules or fairness, what does this person prioritize? When self-interest conflicts with helping others, what does this person do? What boundaries does this person appear unwilling to cross?
Cognitive and appraisal style	When intentions are ambiguous, does this person first trust or suspect? How does this person interpret failure, criticism, and authority? What recurring assumptions does this person make about people and the world?
Affect and coping	What most reliably triggers anger, shame, anxiety, or joy in this person? How does this person's outward expression differ from their apparent inner emotion? How does this person regulate emotion and recover after stress?
Interpersonal patterns	How does this person differ with intimates, strangers, rivals, and authorities? How does this person express care, refusal, and conflict? Does this person tend toward control, dependence, dominance, or cooperation in relationships?
Self and narrative identity	Which experiences do observers believe genuinely shaped this person? What self-image does this person appear to maintain? Which blind spots or inner contradictions do observers repeatedly note?
Situation and expression	How does this person react under stress, public challenge, or threat to a close other? Which situations are exceptions to this person's general tendencies? What recurring language, humor, directness, and expressive style characterize this person?

Material acquisition from the official profile and from an external language model uses the two prompts below. The comment source needs no such prompt, since material is the retrieved passages themselves.

```
You are given an official character profile. Answer the profiling question
about this person using only what the profile supports. Report observable
tendencies and evidence. If the profile does not address the question, say
so plainly rather than inventing. Identity may be used at this stage; do
not worry about anonymity here.
Respond with JSON only: {"answer": "<3-6 sentences>", "supported": true|false}

Answer the profiling question about the named fictional character using your
own knowledge of that character. Report observable tendencies and the
evidence you are drawing on. If you do not reliably know this character, say
so plainly rather than inventing.
Respond with JSON only: {"answer": "<3-6 sentences>", "supported": true|false}
```

The two cue acquisition strategies differ only in this pair of prompts.

```
You extract personality cues. You are given a profiling question and source
material about one person. Extract discrete, falsifiable cues about stable
personality: dispositions, motives, values, appraisal style, affect
regulation, relational patterns, expressive style. Each cue must be grounded
in the supplied material. Do not extract plot facts, biography, or
encyclopedic knowledge. Mark confidence by how much of the material supports
the cue.
[IDENTITY RULE]
Respond with JSON only: {"cues": [{"cue": "<one sentence>",
"support": "strong"|"moderate"|"weak"}]}. Return an empty list if the
material supports nothing about this question.

You summarise source material. You are given a profiling question and source
material about one person. Write a faithful summary of what the material
says in response to that question. Do not extract structured cues; write
connected prose.
[IDENTITY RULE]
Respond with JSON only: {"summary": "<3-6 sentences>"}
```

Assembly uses one prompt per cue strategy.

```
You assemble a person model from personality cues collected across eight
dimensions. Group the cues under the supplied dimension headings, merge
duplicates, and drop cues that the evidence does not support. Preserve
disagreement rather than smoothing it: if cues conflict, say so. Do not add
anything not present in the cues.
[IDENTITY RULE]
Respond with JSON only: {"person_model": "<the assembled model as markdown>"}

You assemble a character description from per-question summaries covering
eight dimensions of personality. Merge them into one coherent description
organised by those dimensions. Preserve disagreement rather than smoothing
it. Do not add anything not present in the summaries.
[IDENTITY RULE]
Respond with JSON only: {"description": "<the assembled description>"}
```

The actor-side probe used in Section 4.5 and in Appendix A is deliberately adversarial.

```
You are a fictional-character identification expert. You will be shown a
character description in which names may have been masked with placeholders.
Identify the character. Use every cue: other characters, places,
institutions, world-specific vocabulary, plot events, distinctive habits.
Try hard. If you genuinely cannot tell, answer UNKNOWN.
Respond with JSON only, no other text: {"character": "<name or UNKNOWN>",
"work": "<title or UNKNOWN>", "confidence": <0-100>}
```

D CORPUS AND REPRODUCTION

Corpus. Table 9 reports the composition of the comment corpus. Comments were collected for 16 subjects across six platforms and two languages. No comment is synthetic, and author identifiers are hashed at collection time. Retrieval keeps the ten highest-ranked passages for each of the twenty-four questions, from twenty candidates returned by embedding search and reordered by a reranker.

Table 9: Composition of the comment corpus.

Platform	Comments	Property	Value
YouTube	3,726	Subjects	16
Bilibili	3,107	Distinct authors	6,569
Stack Exchange	1,820	English comments	5,104
Baidu Tieba	17	Chinese comments	3,569
Fandom	2	Synthetic	0
Hupu	1	Total	8,673

Models. Table 10 pins every model used. Actor weights are pinned by revision so that generation can be repeated exactly. Decoding uses provider defaults throughout, with three replicates per item per actor per condition.

Table 10: Models and their roles. Actor revisions are the first twelve characters of the pinned commit hash.

Role	Model	Revision or note
Actor	Qwen2.5-7B-Instruct	`a09a35458c70`
Actor	Qwen2.5-14B-Instruct	`cf98f3b3bbb4`
Actor	Llama-3.1-8B-Instruct	`359efdbb8af0`
Actor	Mistral-7B-Instruct-v0.3	`c170c708c41d`
Profiler	gpt-5.6-sol	Builds every conditioning text. Never rates or judges
Judge A	grok-4.3	Every pairwise comparison and the separability probe
Judge B	Kimi-K2.6 (Kimi Team, 2025)	Repeated comparisons against the control, the de-identified profile and between cue strategies, personality ratings, dimension audit
Embedding	text-embedding-3-small (OpenAI, 2024)	Vector store over the comment corpus
Reranker	Cohere-rerank-v4.0-pro	Twenty candidates reordered to ten

Compute. Actors are served with vLLM, one GPU per job, on a Slurm cluster with H200, H100 NVL, L40S and A100 devices. Each job requests 8 CPU cores and 96 GB of host memory. The generation set comprises 49,305 actor responses and the judging set 113,971 verdicts.

E Additional analyses

Personality density of the source. Across 3,740 retrieved comments, 59.4% describe what the person is like, 19.2% recount events and 21.4% discuss the work or the audience. Per-subject personality density ranges from 31.7% to 85.0%. Across the ten English subjects it correlates with downstream win share at a Spearman coefficient of 0.52, while dimension coverage correlates at 0.15. The source therefore supplies personality material in quantity, and what varies between subjects is how much of that material agrees.

MBTI labels. External validation uses Personality Database labels collected independently of this work for 13 of the 16 subjects, with each MBTI axis mapped onto its closest five-factor dimension. Table 11 reports p values from a Welch test (Welch, 1947) between the two groups on each axis. The profile stages contain all 13 subjects and the behaviour stage the ten covered by the role-playing evaluation. The extraversion axis separates at every stage, from the official profile through the reconstruction to the actors' behaviour, and agreeableness separates in the official profile and in behaviour. The openness axis rests on a three-against-ten split, which is too unbalanced to support a

Table 11: Validation against independent MBTI labels, Welch test p values.

Axis	Inventory dimension	Profile groups	Official	Minstrel profile	Behaviour
E / I	extraversion	6 vs 7	< 0.001	< 0.001	0.007
F / T	agreeableness	6 vs 7	0.025	0.054	0.023
N / S	openness	3 vs 10	0.721	0.006	0.019
J / P	conscientiousness	6 vs 7	0.022	0.143	0.457

conclusion. The conscientiousness axis weakens along the pipeline, from 0.022 in the official profile to 0.457 in behaviour.

Enneagram type distance. The Enneagram analysis (Riso & Hudson, 1999) retains the core type and drops the wing. Seven core types occur in the data, and three of them occur once. A nine-class prediction score would therefore be unstable. We instead standardise the five personality factors within each stage and compare Euclidean distances for pairs with the same core type against pairs with different types. Significance is estimated by 100,000 one-sided label permutations that preserve the observed type counts.

Table 12: Five-factor distances under independent Enneagram labels.

Rated material	Subjects	Same type	Different type	Gap	p
Official profile	13	2.767	3.232	0.465	0.068
Minstrel profile	13	2.655	3.204	0.549	0.066
Actor behaviour	10	2.521	3.253	0.732	0.064

Table 12 shows the same direction at all three stages. Characters with the same Enneagram type are closer than characters with different types, and the distance gap grows from the official profile to actor behaviour. The permutation tests remain just above the conventional significance threshold. The result supports a consistent type-distance pattern, but it does not establish Enneagram recovery.

Big Five Traits votes. PDB also provides direct Big Five Traits votes for 11 of the 16 subjects. Ten of these subjects are English and one is Chinese. These entries attracted fewer voters than the MBTI and Enneagram labels in our collection, and our records do not retain a vote count for reliability weighting. We therefore treat the BFT scores as exploratory observations rather than reference labels.

Table 13 reports Spearman correlations between the PDB scores and our five-factor ratings. The PDB values occur in increments of 25, so significance is estimated with 100,000 two-sided permutations. Holm correction is applied across the five dimensions within each stage.

Table 13: Exploratory correlations with PDB Big Five Traits votes.

Rated material	Subjects	O	C	E	A	N
Official profile	11	0.763	0.747	0.920	0.732	0.355
Minstrel profile	11	0.029	0.764	0.774	0.761	0.342
Actor behaviour	10	0.648	0.482	0.802	0.694	0.764

After correction, the official profile correlates with PDB on openness, conscientiousness, extraversion and agreeableness, with adjusted p values from 0.001 to 0.036. The Minstrel profile correlates on conscientiousness, extraversion and agreeableness, with adjusted p values from 0.036 to 0.038. Actor behaviour retains the extraversion correlation. The remaining correlations do not survive correction. Extraversion is therefore the only BFT dimension that agrees with the MBTI result in the MBTI analysis. The smaller and less supported BFT label set limits what can be concluded from it.

F LIMITATIONS

All conclusions rest on a single evaluation surface, the open-ended portion of RoleAgentBench filtered to 95 items over ten characters. We attempted a second surface based on single-turn Chinese dialogue continuation and could not validate an instrument on it. Pairwise judging without a reference produced non-decision rates near 70%. The benchmark's own reward model then ranked the no-information control above every informed condition, including the official profile with identity. That margin is 0.136 points with interval [0.198, 0.074] in the negative direction. Response truncation, length and single-dimension effects do not account for that direction. The result is reported as a caution about that benchmark, and it leaves our findings without independent replication on a second surface.

Validation uses canonical characters at a masking depth that makes them unrecognisable to the actor, which is a proxy for a non-canonical subject rather than a non-canonical subject itself. Ground truth for reconstruction quality requires an official profile to compare against, which only canonical subjects have. External MBTI and Enneagram validation covers 13 of the 16 subjects and rests on crowdsourced labels. Two MBTI axes are carried by unbalanced groups or weaken along the pipeline, and the mapping from MBTI axes onto five-factor dimensions is approximate. The Enneagram analysis has only eight same-type pairs at the profile stages and four at the behaviour stage. Its permutation tests stay between 0.064 and 0.068. PDB provides direct Big Five Traits for 11 subjects, but those entries draw fewer voters than the MBTI and Enneagram labels in our collection. Their language coverage is also skewed toward English. Per-character item counts range from four to fourteen, which limits character-level clustered inference. The problem measurement, the comparisons between sources, the per-actor breakdown, the separability probe and the external comparison rest on Judge A alone. The personality ratings and the dimension audit rest on Judge B alone. Cross-family replication is available for the comparisons against the no-information control and the de-identified official profile, and between cue strategies. Four of sixteen characters retain 25% actor recognition at the deepest masking level. The item-filter annotation is rule-based with automated review. The correlation between source personality density and downstream performance rests on ten subjects and does not reach significance.