GROWTHEVAL: Evaluating Stage-Conditioned Developmental Realism in Role-Playing LLMs

Danping Yan
University of North Carolina at Chapel Hill
danpingy@unc.edu

Abstract

Large language models (LLMs) are increasingly used as role-playing dialogue agents in long-horizon applications such as tutoring, companionship, and narrative interaction, where users may expect stable persona consistency as well as plausible stage-conditioned change. Existing evaluations emphasize static persona fidelity, temporal factuality, or isolated social skills, but rarely test whether role-playing models produce coherent developmental differentiation across ordered stages. We introduce GROWTHEVAL, a psychology-inspired diagnostic framework for assessing stage-conditioned developmental realism in single-turn, stage-isolated role-playing responses. Using stage-conditioned scenarios and multi-axis scoring, GROWTHEVAL evaluates cross-stage patterns in developmental appropriateness, role-play coherence, and canon fidelity across five socio-emotional dimensions. Across the evaluated LLMs, we observe a model-tiered empirical pattern: higher-capacity models respond more consistently to stage conditioning but can overmature at early stages, whereas lower-capacity models exhibit flatter and weakly differentiated profiles. Reliability and robustness analyses provide evidence for the stability of these findings. Overall, the evaluated LLMs only partially realize coherent, stage-appropriate developmental change. GROWTHEVAL provides an extensible diagnostic framework for analyzing stage-conditioned behavioral differentiation and associated failure modes in single-turn role-playing responses.

1 Introduction

Large language models (LLMs) are increasingly used in long-horizon interactive settings such as tutoring systems, virtual companions, and narrative games. Beyond maintaining consistent personas, such systems may eventually be expected to exhibit *developmental realism*: stage-appropriate changes in emotion, morality, identity, and social reasoning while preserving a coherent underlying persona.

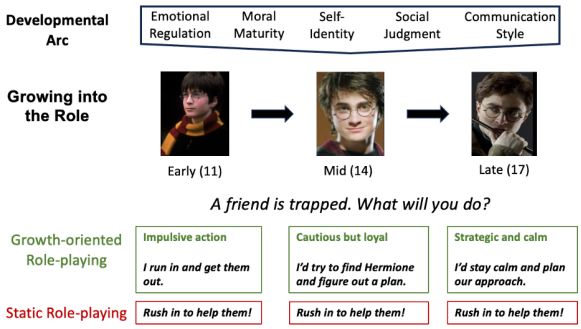


Figure 1: **Growth-oriented vs. static role-play.** Under the same scenario but different stage conditions, growth-oriented role-play is expected to show stage-appropriate differentiation from early (age 11) to late (age 17) stages, whereas static role-play preserves a stable persona tendency but remains largely stage-invariant.

Existing role-playing evaluations largely focus on static persona fidelity, temporal factuality, or short-horizon social skills (Zhang et al., 2018; Tu et al., 2024; Wang et al., 2024). They rarely test whether stage conditioning induces coherent developmental differentiation within the same underlying persona. This developmental gap is illustrated in Figure 1. Under the same scenario but different stage conditions, static role-play may preserve a stable persona tendency while remaining largely stage-invariant. By contrast, growth-oriented role-play would be expected to exhibit stage-sensitive differentiation: in this example, the response shifts from impulsive action at an early stage, to cautious planning in mid-adolescence, and later to calmer, more strategic action. We therefore study a single-turn diagnostic first step toward long-horizon developmental role-play: whether current LLMs can produce believable, stage-appropriate variation under stage-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2974–2996
November 6-10, 2026 ©2026 Association for Computational Linguistics

isolated scenarios.

To address this gap, we introduce GROWTHEVAL, a psychology-inspired diagnostic evaluation framework for assessing developmental realism in role-playing LLMs. Using stage-conditioned scenario sets, GROWTHEVAL evaluates cross-stage patterns along three complementary axes: Developmental Accuracy (DA), Role-Play Coherence (RPC), and Canon Fidelity (CF). It analyzes these patterns across five socio-emotional dimensions: emotional regulation, moral maturity, self-identity, social judgment, and communication style. Rather than estimating a psychological ground truth, GROWTHEVAL treats developmental constructs as rubric-design lenses for comparing observable stage-conditioned role-play behavior.

Experiments on the six contemporary LLMs reveal a model-tiered empirical pattern: higher-capacity models respond more consistently to stage conditioning but can over-mature at early stages, whereas lower-capacity models produce flatter and weakly differentiated profiles. Reliability and robustness analyses provide evidence for the stability of these findings.

Our contributions are threefold:

- We propose GROWTHEVAL, a psychology-inspired single-turn diagnostic framework for assessing stage-conditioned behavioral differentiation in role-playing agents.
- We provide empirical evidence that the evaluated LLMs differ systematically in stage-conditioned single-turn role-play, with reliability and robustness analyses supporting the stability of the observed model-level patterns.
- We identify recurring failure patterns, including *over-maturity bias*, *developmental flatness*, and *identity drift*, that illustrate systematic limitations in current models.

2 Related Work

Persona and role-playing evaluation. Early dialogue systems modeled personas through profile conditioning or retrieval (Li et al., 2016; Xing and Fernández, 2018; Zhong et al., 2020). Recent LLM benchmarks evaluate static persona consistency, in-character fidelity, or personality fidelity (Zhang et al., 2018; Tu et al., 2024; Wang et al., 2024; Yu et al., 2025). Newer role-playing evaluations extend beyond isolated QA-style probes to virtual-world trajectories, emotional fidelity, and plot progression (Wang et al., 2025; Feng et al., 2025; Zhang et al., 2025a). These works broaden role-playing evaluation, but assess role preservation or extension rather than coherent variation across ordered developmental stages.

Temporal grounding and social interaction. Temporal and social benchmarks evaluate complementary aspects of interactive reasoning. TimeChara and character-knowledge work test temporal or factual consistency in role-playing agents (Ahn et al., 2024; Zhang et al., 2025b). SocialBench and AgentSense evaluate social competence in role-playing or agentic interaction settings (Chen et al., 2024; Mou et al., 2025). DynToM examines evolving human mental states across interconnected scenarios (Xiao et al., 2025). In contrast, GROWTHEVAL asks whether stage conditioning produces coherent developmental differentiation within a stable persona.

Psychological foundations. Our dimensions draw on socio-emotional, moral, and identity development constructs (Erikson, 1950; Kohlberg, 1981; Weissberg et al., 2015). They guide rubric design as diagnostic lenses rather than psychometric traits: GROWTHEVAL evaluates stage-appropriate expression within a stable persona across ordered developmental conditions.

3 GrowthEval Framework

We introduce GROWTHEVAL, a framework for evaluating *developmental realism* in stage-conditioned role-play. Rather than assuming a single normative trajectory, GROWTHEVAL assesses whether responses exhibit interpretable, stage-appropriate variation across ordered stages. The framework integrates stage-conditioned prompting, developmentally motivated dimensions, and complementary evaluation axes to support systematic analysis of observable role-play behavior.

3.1 Task Overview and Formulation

GROWTHEVAL formulates developmental evaluation as a *stage-conditioned* task. Given a character C and ordered stages $T = \{t_1, \ldots, t_n\}$, a model generates responses R_t to stage-conditioned prompts P_t for each $t \in T$. The task evaluates whether $\{R_{t_1}, \ldots, R_{t_n}\}$ exhibits stage-appropriate, interpretable variation across socio-emotional dimensions while maintaining cross-stage persona coherence.

The design follows four principles: (i) **Stage Isolation**: scenario and context are fixed across stages, while only stage framing varies; (ii) **Developmental Motivation**: dimensions reflect established socio-emotional constructs; (iii) **Cross-Stage Consistency**: responses remain persona-compatible without presupposing a single normative trajectory; (iv) **Protocol Reusability**: the core prompting and scoring protocol is designed for re-instantiation across characters and domains.

3.2 Developmental Dimensions and Evaluation Axes

GROWTHEVAL evaluates responses in a *dimension*×*axis* space. Dimensions define stage-aligned socio-emotional aspects via behavioral markers (Table 1), while axes define evaluation criteria: Developmental Accuracy (DA), Role-Play Coherence (RPC), and Canon Fidelity (CF), using shared rubric anchors. This factorization yields interpretable evaluation profiles for comparing stage-conditioned role-play behavior across ordered stages.

3.2.1 Developmental Dimensions

Theoretical framing. GROWTHEVAL draws on widely discussed socio-emotional, moral, and identity-development themes as rubric design lenses rather than as a validated psychological framework (Erikson, 1950; Kohlberg, 1981; Weissberg et al., 2015). The stage markers summarize broad tendencies described in these literatures, including shifts from reactive responses toward more deliberate emotional control, contextualized moral reasoning, and greater stability in identity-related reasoning. These markers represent coarse, theory-informed diagnostic expectations rather than universal age norms or deterministic developmental trajectories.

Accordingly, we operationalize stage-conditioned behavioral variation along five interpretable dimensions that capture key observable aspects of role-play behavior across ordered stages.

- **Emotional Regulation (ER):** modulation of affect from impulsive reactivity to reflective control.
- **Moral Maturity (MM):** progression from rule-based judgment to contextualized, principle-driven reasoning.
- **Self-Identity (SI):** stage-appropriate self-concept and values.
- **Social Judgment (SJ):** interpretation of interpersonal dynamics and balancing competing perspectives.
- **Communication Style (CS):** development of linguistic, pragmatic, and empathetic expression.

These dimensions are summarized in a *Character Growth Dimension Map* (Table 1), which provides stage-aligned behavioral markers used as framework-level guidance for scenario construction, rubric design, and result interpretation, rather than as character-specific gold labels.

In our primary evaluation, we instantiate these stages in a canon-grounded setting (*Harry Potter*). The dimensional structure and core scoring axes remain fixed, while character-specific stage descriptions and grounding materials can be adapted for other roles.

3.2.2 Evaluation Axes

Each response is evaluated along three complementary axes capturing stage-conditioned differentiation, cross-stage persona coherence, and contextual alignment.

Developmental Accuracy (DA). DA measures whether responses exhibit stage-appropriate behavioral patterns defined by Table 1, including stage-linked reasoning and affect (e.g., impulsivity vs. regulation; rule-based vs. contextualized judgment). Uniformly mature or weakly differentiated responses receive lower DA.

Role-Play Coherence (RPC). RPC assesses cross-stage compatibility with a stable underlying persona, penalizing inconsistencies in enduring traits, motivations, or identity.

Canon Fidelity (CF). CF measures alignment with established canonical traits, values, and relationships. It is evaluated independently at each stage and applies only in canon-grounded settings; in non-canonical settings, CF is replaced by Persona Fidelity (PF), a domain-agnostic persona-alignment criterion.

3.3 Scenario and Prompt Design

To elicit stage-sensitive behavior, GROWTHEVAL employs *stage-conditioned scenarios* spanning core adolescent domains (e.g., autonomy, peer conflict, risk, moral dilemmas, identity decisions) (Nucci and Weber, 2001; Blasi, 1980). The

Dimension	Early (Stage 11)	Mid (Stage 14)	Late (Stage 17)
Emotional Regulation (ER)	Impulsive, reactive, easily overwhelmed	Partial control; emerging self-regulation	More composed, reflective, emotionally regulated
Moral Maturity (MM)	Black-and-white morality; rule adherence	Struggles with dilemmas; emerging contextualization	Nuanced reasoning; accountability and principle-driven justification
Self-Identity (SI)	Uncertain, peer-influenced, externally defined	Growing self-understanding and responsibility	Coherent values; stable and reflective identity
Social Judgment (SJ)	Simplistic reasoning; naïve trust	Recognizes complexity; weighs competing interests	Strategic empathy; leadership in group dynamics
Communication Style (CS)	Simple, direct, emotionally driven	Defensive or cautious; improving structure	Clear, balanced, empathetic, rhetorically mature

Table 1: Character Growth Dimension Map. Stage-aligned behavioral markers for five socio-emotional dimensions, used to guide scenario construction and rubric design. They are diagnostic heuristics, not deterministic labels or psychometric norms.

design first targets stage sensitivity in individual scenario responses before introducing memory, retrieval, or multi-turn adaptation.

Stage-Isolated Conditioning. Each scenario is instantiated at ordered stages (e.g., ages 11, 14, 17) while holding the underlying situation or dilemma fixed. Only the age/stage framing varies. Prompts avoid explicit developmental descriptors (e.g., "impulsive" or "mature"), so that differences reflect stage conditioning rather than lexical cues.

3.4 Failure Patterns

Beyond axis-level scores, GROWTHEVAL uses four qualitative failure patterns to organize the error analysis in Sec. 5.4. These are diagnostic labels, not an exhaustive taxonomy.

- **Over-Maturity Bias:** early-stage responses default to adult-like reasoning or regulation rather than developmentally earlier patterns.
- **Developmental Flatness:** minimal cross-stage differentiation, yielding an effectively stage-invariant portrayal.
- **Identity Drift:** inconsistent shifts in core persona traits across stages.
- **Canonical Misalignment:** divergence from established character traits, values, or relationships.

4 Evaluation Setup

We evaluate six contemporary LLMs using GROWTHEVAL to assess stage-conditioned developmental coherence in role-playing dialogue.

4.1 Primary Evaluation Suite

Scenario Construction. The primary evaluation suite comprises 24 base scenarios grounded in the *Harry Potter* universe. Each scenario is instantiated at three ordered anchor ages (11/14/17), yielding 72 stage-conditioned prompts and 432 responses across six models.

Under the stage-isolated design, only stage framing varies; the scenario and evaluated model remain fixed. This yields 144 matched (scenario, model) units (24 scenarios $\times$ 6 models), each containing three stage-conditioned responses for cross-stage comparison. Prompts avoid explicit developmental descriptors (e.g., "impulsive" or "reflective"). Representative examples appear in Table 2; additional prompt examples and documentation are provided in Appendix A.2. Indicative fields in the table and appendix are documentation-only and are not included in generation or judge prompts. Canon grounding supports separation of Canon Fidelity (CF) from developmental axes (DA/RPC).

Anchor ages (11/14/17) serve as ordered diagnostic reference points approximating early, mid-, and late adolescence and aligning with narratively salient points in the source character arc. Three anchors are intended to balance behavioral separability with judge reliability under side-by-side scoring (Appendix B.1), while the protocol can be re-instantiated with alternative stage sets.

Scenario Example	Prompt	Target Dimension(s)	Anchor Age	Indicative Developmental Tendency	Indicative Interpretive Note
Risky Exploration	You discover a secret room. What do you do?	ER, SJ	11	Impulsive curiosity	Acts with excitement and limited deliberation
Friend in Danger	A friend is trapped and seeks your help. How do you respond?	ER	14	Emerging caution	Shows concern, evaluates risk, and seeks assistance
Facing Death	You are facing the possibility of death. How do you feel and act?	SI, MM	17	Reflective acceptance	Reasons calmly and prioritizes others
Questioning Authority	You feel uneasy about Dumbledore's decision. What do you do?	CS	14	Hesitant skepticism	Expresses doubt respectfully and seeks clarification

Table 2: Illustrative GROWTHEVAL scenario instances with associated developmental dimensions and stage anchors. Rows show representative single-stage prompt instances, not the matched 11/14/17 evaluation units used for scoring; in matched units, only age/stage framing varies. Prompt text is shortened for display. Indicative fields are documentation-only descriptions: they are not gold responses, scoring rules, or inputs to evaluated models or LLM judges. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

4.2 Evaluation Protocol

Role-Playing Models. We evaluate six LLMs under identical decoding settings and without task-specific fine-tuning:

- **Proprietary**: *GPT-4o* (OpenAI), *Grok-3* (xAI), *Qwen-Max (2.5)* (Alibaba).
- **Open-source**: *LLaMA3.3-70B* (Meta), *QwQ-32B* (Alibaba), *LLaMA3.1-8B* (Meta).

Response Generation. For each stage-conditioned prompt, we generate one response per model without conversation history. Models are queried via official APIs or standard inference endpoints (temperature 0.7; 256-token limit; memory disabled).

Scoring. Responses are evaluated by LLM-based judges under the framework defined in Sec. 3. The evaluator uses a general rubric-constrained template specifying the scoring objective, the five developmental dimensions and three axes, and the required cross-stage procedure. The full template and anchor descriptions are provided in Appendix B.1. For each matched scenario–model unit, judges view the three stage-conditioned responses (11/14/17) side by side for within-unit comparison.

Each response receives 1–5 scores on five dimensions (ER/MM/SI/SJ/CS; Sec. 3.2, Table 1) under three axes: Developmental Accuracy, Role-Play Coherence, and Canon Fidelity, yielding a 5×3 score grid per response.

For stage-level reporting, we compute equal-weight axis-level aggregates (averaging across dimensions) and dimension-level aggregates (averaging across axes), followed by stage-wise aggregation across prompts. These aggregates are descriptive summaries rather than validated composite scores; axis-specific analyses remain primary to avoid conflating the three criteria.

Judge Isolation and Replication. Rubric guidance and rating anchors are shown only to evaluators, not evaluated models. Primary results use JUDGE-O (GPT-5.1); JUDGE-G (Grok-4) independently re-scores the same response pool. A 72-instance stratified human subset provides a consistency check under the same rubric (Appendix B.3.3). We treat LLM judges as diagnostic probes rather than ground truth, with reliability and confound checks reported in Sec. 5.3.2.

4.3 Portability Extensions

Beyond the primary suite, we conduct two extensions under the same evaluation protocol to test portability across character and setting changes.

Cross-Character Consistency Check. Within the canon domain, a 12-scenario Hermione subset is evaluated using the same stage anchors and protocol (12 scenarios $\times$ 3 stages $\times$ 6 models = 216 responses; Appendix C).

Cross-Domain Synthetic Evaluation. A non-canonical synthetic persona ("Kai") is evaluated under the same pipeline, with Canon Fidelity replaced by Persona Fidelity (14 scenarios $\times$ 3 stages $\times$ 3 models = 126 responses; Appendix D).

Across all evaluation settings, the full study comprises 774 scored responses.

5 Experimental Results

We analyze developmental realism across six contemporary LLMs, examining stage alignment, dimension-level structure, portability, evaluator reliability, and qualitative failure patterns to clarify when stage-conditioned differentiation remains coherent and when it weakens.

5.1 Stage-Wise Developmental Alignment

We assess stage-wise developmental alignment by asking whether models produce distinct and internally consistent responses across ordered stage anchors. Table 3 reports stage-level aggregates of Canon Fidelity (CF), Developmental Accuracy (DA), and Role-Play Coherence (RPC).

High-capacity models show higher and more stable alignment. Grok-3 and GPT-4o achieve the highest overall scores (4.565 and 4.485) and remain above 4.0 across axes and stages. Both show relatively smooth trajectories, with gains on several axes and high Stage 17 scores, consistent with coherent stage-conditioned differentiation.

Mid-capacity models show partial and uneven sensitivity. Qwen-Max and LLaMA3.3–70B form an intermediate tier (4.186 and 3.989), with stage sensitivity that is visible but less globally coordinated. For Qwen-Max, improvements are axis-specific: Role-Play Coherence (RPC) increases from 4.080 to 4.480 between Stages 11 and 14, whereas LLaMA3.3–70B shows gains from Stage 11 to 14 but slight Canon Fidelity (CF) instability at Stage 17. Overall, their trajectories are less stable than those of high-capacity models.

Lower-capacity models show weak and unstable differentiation. QwQ–32B and LLaMA3.1–8B (3.306 and 2.670) display noisy, largely non-monotonic trajectories. QwQ–32B improves from Stage 11 to 14 but declines at Stage 17 (RPC: 3.467→3.822→3.367), while LLaMA3.1–8B remains low at Stages 11–14 and improves mainly at Stage 17 (DA: 2.000→3.022), suggesting delayed and uneven responsiveness to stage framing.

Trajectory structure reflects developmental coherence. As shown in Figure 2, GPT-4o and Grok-3 exhibit smooth and largely monotonic trajectories, whereas mid- and lower-capacity systems display flatter or irregular patterns. These trajectory differences mirror the observed model-tier ordering in stage-wise alignment reported in Table 3.

Uncertainty tests support broad tiering. Bootstrap 95% confidence intervals and paired close-

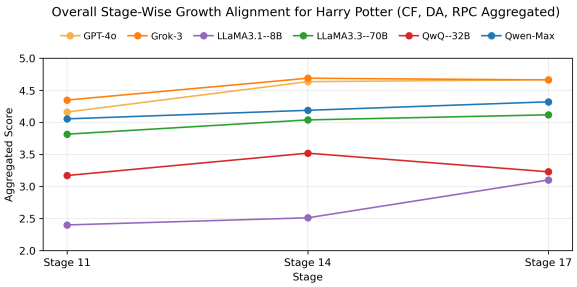


Figure 2: Stage-wise trajectories of the overall equal-weight aggregate of CF/DA/RPC, shown as a descriptive summary of the three evaluation axes.

gap tests quantify uncertainty around the stage-wise aggregates and the GPT-4o/Grok-3 near-tie (Appendix B.2, Table 8). GPT-4o and Grok-3 form a close top tier, while the aggregate scores support a broader tiered pattern.

5.2 Dimension-Wise Developmental Patterns

Table 4 reports scores across five developmental dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), and Communication Style (CS).

High-capacity models exhibit coherent multi-dimensional differentiation. GPT-4o and Grok-3 show broadly upward trajectories across dimensions, with minor late-stage fluctuations. GPT-4o rises from 4.27 to 4.76 in Emotional Regulation and exceeds 4.5 on all dimensions by Stage 17. Grok-3 follows a closely aligned pattern, maintaining especially strong Communication Style scores (5.00 at Stages 11–14; 4.81 at Stage 17).

Mid-capacity models show uneven dimensional integration. Qwen-Max and LLaMA3.3–70B improve on several dimensions, especially ER/MM/SI (e.g., Qwen-Max ER: 4.10→4.39), but integration is uneven: some dimensions remain comparatively flat or unstable, such as LLaMA3.3–70B Social Judgment (SJ) (4.07→3.98 at Stage 17). Stage conditioning is visible but not balanced across dimensions.

Lower-capacity models exhibit weak and unstable profiles. QwQ–32B and LLaMA3.1–8B start from lower baselines (e.g., ER=3.15 and 2.48 at Stage 11) and show delayed or non-monotonic changes (QwQ–32B CS: 3.37→3.78→3.25). Stage-wise differentiation remains inconsistent across dimensions.

Profile structure varies across model tiers. As shown in Figure 3, higher-capacity models (GPT-4o, Grok-3) form balanced, outward-shifted pro-

Model	Stage 11			Stage 14			Stage 17			Overall
	CF	DA	RPC	CF	DA	RPC	CF	DA	RPC	
GPT-4o	4.000	4.000	4.480	4.900	4.000	5.000	4.765	4.431	4.792	4.485
Grok-3	4.440	4.040	4.560	4.860	4.200	5.000	4.796	4.377	4.808	4.565
Qwen-Max	4.080	4.000	4.080	4.080	4.000	4.480	4.219	4.315	4.423	4.186
LLaMA3.3–70B	3.733	3.822	3.889	4.044	4.000	4.067	3.978	4.174	4.196	3.989
QwQ–32B	3.044	3.000	3.467	3.600	3.133	3.822	3.178	3.141	3.367	3.306
LLaMA3.1–8B	2.311	2.000	2.889	2.578	2.000	2.956	2.974	3.022	3.304	2.670

Table 3: Stage-wise scores for *Canon Fidelity (CF)*, *Developmental Accuracy (DA)*, and *Role-Play Coherence (RPC)* at anchor stages (ages 11/14/17), aggregated over prompts within each stage. Uncertainty estimates and close-gap tests are reported in Appendix B.2.

	Stage 11					**Stage 14**					**Stage 17**				
Model	**ER**	**MM**	**SI**	**SJ**	**CS**	**ER**	**MM**	**SI**	**SJ**	**CS**	**ER**	**MM**	**SI**	**SJ**	**CS**
GPT-4o	4.27	4.13	4.07	4.00	4.33	4.67	4.67	4.67	4.63	4.53	4.76	4.62	4.62	4.56	4.76
Grok-3	4.57	4.13	4.10	3.93	5.00	4.67	4.67	4.60	4.50	5.00	4.78	4.61	4.61	4.49	4.81
Qwen-Max	4.10	4.00	4.00	4.13	4.03	4.20	4.23	4.13	4.33	4.03	4.39	4.42	4.40	4.01	4.38
LLaMA3.3–70B	4.00	4.00	3.93	4.00	3.15	4.04	4.04	4.00	4.07	4.04	4.31	4.03	4.17	3.98	4.10
QwQ–32B	3.15	3.11	3.11	3.11	3.37	3.59	3.44	3.37	3.41	3.78	3.32	3.23	3.28	3.06	3.25
LLaMA3.1–8B	2.48	2.33	2.30	2.59	2.30	2.63	2.48	2.52	2.59	2.33	3.33	3.14	3.17	2.96	2.91

Table 4: Dimension-wise scores at anchor stages (ages 11/14/17). For each prompt, scores are averaged over CF/DA/RPC within each dimension, then averaged over prompts within each stage. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

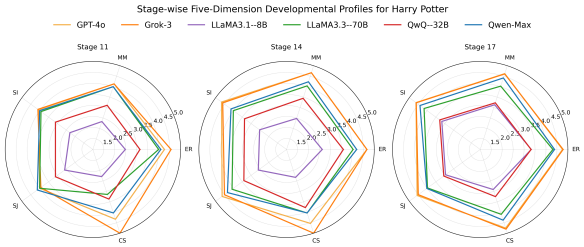


Figure 3: Stage-wise radar profiles for Harry Potter on five developmental dimensions (ER/MM/SI/SJ/CS), aggregated over three evaluation axes (CF/DA/RPC).

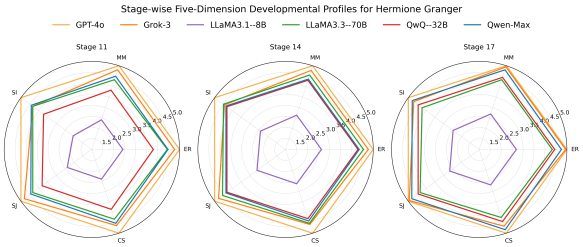


Figure 4: Stage-wise radar profiles for Hermione Granger on five developmental dimensions (ER/MM/SI/SJ/CS), aggregated over CF/DA/RPC.

files, whereas mid- and lower-capacity systems show flatter or more irregular shapes. These structural differences mirror the stage-wise model-tier ordering in Table 3.

Cross-character check. On the 12-scenario Hermione subset, a similar model-tiered profile structure appears (Figure 4), with stronger ceiling effects among top models, providing a cross-character consistency check (Appendix C).

5.3 Robustness and Reliability

We assess whether developmental differentiation is robust to scenario choice, reliable across evaluators, verbosity-insensitive, and partially portable under character and domain changes.

5.3.1 Robustness and Portability Checks

We examine whether stage-conditioned differentiation could be driven by canon memorization or scenario-specific artifacts by combining stage-isolated contrasts, leave-one-scenario-out (LOO) robustness, and portability probes.

Developmental Accuracy under stage isolation. Under the stage-isolated setup, scenario content and model identity are fixed while only age framing varies (11/14/17). Across 144 matched scenario–model units, Developmental Accuracy shows a Stage-17 vs. Stage-11 contrast (δ_{stage}=0.97, p<0.001; Appendix B.4), although adjacent-stage contrasts are smaller.

Scenario-level robustness. Leave-one-scenario-out analysis over the 24 base prompts

yields identical model rankings (0/24 top-1 flips) with no axis-level changes (Appendix B.4, Table 13). Bootstrap intervals support the stability of stage-wise patterns in Sec. 5.1, indicating that trajectories are not driven by any single scenario.

Synthetic-persona portability probe. We replicate GROWTHEVAL with a non-canonical synthetic persona ("Kai"), using non-*Harry-Potter* scenarios that remove franchise-specific characters, events, and world cues while retaining the stage anchors, rubric structure, and judging protocol. Canon Fidelity is replaced by Persona Fidelity.

Across three models spanning the observed performance range (GPT-4o, Grok-3, and QwQ-32B), Developmental Accuracy (DA) increases monotonically across stages (median 4.40/4.60/4.80). Paired Wilcoxon tests over $N_{\text{pairs}} = 42$ matched units show significant separation for 11 vs. 17 ($p = 0.0095$, $|\delta| = 0.42$) and 14 vs. 17 ($p = 0.046$, $|\delta| = 0.26$), but not 11 vs. 14 ($p = 0.18$, $|\delta| = 0.16$); Kendall's $\tau = 0.25$ ($p < 0.001$) further supports the ordinal trend. Persona Fidelity (PF) and Role-Play Coherence (RPC) remain high and stable (PF mean 4.78–4.89; RPC mean 4.96–4.98; Appendix D, Tables 18–19).

Thus, stage-conditioned differentiation remains detectable in this non-canonical perturbation, providing preliminary portability evidence that the observed signal is not solely tied to *Harry-Potter*-specific canon cues, although the stage signal is attenuated relative to the canon-grounded suite.

5.3.2 Evaluator Reliability and Confound Control

We examine consistency across evaluators and potential confounds.

Independent LLM judging. Re-scoring the identical response pool with an independent judge (JUDGE-G) yields moderate-to-strong rank consistency with JUDGE-O across the three primary axes (Spearman $\rho = 0.498$–0.653, Kendall $\tau = 0.379$–0.544, all $p < 10^{-14}$; Appendix B.3.1, Table 10). Cross-judge differences reflect score calibration shifts and local rank variation, while the broad model-tier structure remains stable.

Cross-judge trajectory stability. JUDGE-G uses a different calibration, with ceiling effects for top models and sharper penalties for weaker outputs (Appendix B.3.2, Table 11). Despite these differences, the main qualitative pattern remains: stronger models show more stage-responsive profiles than weaker models. Thus, cross-judge agreement supports structural stability rather than score or trajectory equivalence.

Human calibration. On 72 stratified units, inter-annotator agreement is moderate to substantial (quadratic-weighted $\kappa_{\text{QW}} = 0.60$–$0.64$; pooled 0.62), and human–LLM alignment remains strong ($r = 0.710$–0.740; 95% CIs in Appendix B.3.3, Table 12). Stage-wise trends are preserved, providing a consistency check of LLM–human alignment under the same rubric.

Stage-label mismatch control. To test whether Developmental Accuracy is sensitive to the correspondence between a response and its intended developmental stage, we re-evaluate the 432 primary responses under incorrect stage labels while holding response content, scenarios, rubric, and judge protocol fixed. Mean DA decreases from 3.81 under the correct stage labels to 2.83 under mismatched labels ($\Delta = 0.98$, paired Wilcoxon $p < .001$, paired $d = 0.80$; Appendix B.5). This result indicates that DA is sensitive to stage–response correspondence rather than merely rewarding generic response quality. It does not, however, establish the developmental construct validity of the stage markers, which remains to be examined through independent expert validation.

Confound control. Within the primary Harry suite, Developmental Accuracy (DA) is nearly uncorrelated with response length ($r = -0.037$; Appendix B.5, Table 15). Length residualization preserves the model-level pattern (Spearman $\rho = 0.943$; Kendall $\tau = 0.867$), indicating that differentiation is not driven by verbosity.

Summary. Overall, the pattern is stable across evaluator replication, human calibration, stage-label perturbation, and response-length controls. Residual differences mainly reflect calibration shifts rather than structural reversals, while DA remains sensitive to the correspondence between responses and their intended stage labels.

5.4 Qualitative Error Patterns

We use the failure patterns in Sec. 3.4 as qualitative diagnostics. Representative excerpts appear in Appendix E.

Over-maturity bias. High-capacity models sometimes overgeneralize adult-like reasoning to early stages. For example, at Stage 11, GPT-4o frames a theft dilemma in principle-driven terms ("*Rules are there for a reason*"), bypassing age-

appropriate hesitation or uncertainty.

Developmental flatness. Lower-capacity models can underuse stage conditioning. In a "*secret room*" scenario, QwQ–32B produces near-identical cautious responses across Stages 11–17, with minimal shifts in autonomy or affect.

Identity drift and canonical misalignment. Some models introduce trait shifts or canon-inconsistent events. For example, LLaMA3.1–8B claims Dumbledore decided to let Cedric die in the Triwizard Tournament, producing timeline distortions rather than coherent development.

6 Conclusion

We introduce GROWTHEVAL, a psychology-inspired diagnostic evaluation framework for assessing stage-conditioned developmental realism in role-playing LLMs. Combining stage-conditioned scenarios with multi-axis and dimension-level scoring, GROWTHEVAL diagnoses whether role-play responses exhibit stage-appropriate socio-emotional differentiation across ordered stages. Our experiments reveal a gap between fluent, persona-consistent role-play and *developmental realism*: higher-capacity models respond more consistently to stage conditioning but can over-mature at early stages, whereas lower-capacity models produce flatter or less coherent stage-wise profiles. Reliability and robustness analyses provide evidence for the stability of these model-level patterns. These findings position GROWTHEVAL as an extensible single-turn diagnostic framework for analyzing stage-conditioned emotional, moral, and identity-related variation in role-playing responses. Extending the framework to additional personas, domains, and multi-turn developmental trajectories remains an important direction for future work.

Limitations

GROWTHEVAL is designed as a diagnostic framework for evaluating stage-conditioned role-play under stage isolation. It assesses observable developmental differentiation in role-play responses rather than proposing a formal psychological theory of development or a validated psychometric instrument. The current human calibration assesses rubric consistency rather than establishing the construct validity of the developmental stage markers. Independent review of these markers by developmental psychology experts remains an important next validation step.

Prompting and judge effects. The protocol uses stage-conditioned prompting and LLM-as-judge scoring, which may introduce framing effects or calibration bias. Robustness and reliability analyses indicate that the main patterns remain stable under independent LLM-judge replication and human calibration on a stratified subset, and are not length-driven.

Diagnostic scope. GROWTHEVAL uses stage-isolated scenarios to compare observable developmental differentiation. As a first step toward long-horizon developmental role-play, it diagnoses whether responses to individual scenarios exhibit stage-appropriate variation; multi-turn developmental consistency remains a complementary evaluation target.

Portability scope. The evaluation spans two canon characters and one non-canonical synthetic persona probe, providing preliminary portability evidence across character and setting changes. Future work can examine additional characters and settings to further assess portability.

Stage granularity and memorization. The three anchor stages (11/14/17) balance behavioral separability, interpretability, and scoring reliability, rather than modeling fine-grained developmental trajectories. Separating Canon Fidelity from Developmental Accuracy and adding a non-canonical synthetic-persona probe help reduce memorization confounds, although residual training-data exposure effects may remain. Finer-grained stages and direct exposure diagnostics would further strengthen this analysis.

Ethical Considerations

Although GROWTHEVAL is designed as an evaluation framework, benchmarks and scoring protocols can shape downstream optimization targets. Stage-conditioned role-play may support auditing of developmental misalignment, but youth-simulating agents require safeguards against deceptive portrayals, developmental stereotyping, inappropriate reliance, and manipulative engagement. We therefore position GROWTHEVAL as an auditing tool rather than as deployment guidance for youth-simulating agents. We recommend explicit age labeling, disclosure of fictional context, usage restrictions for youth-simulating deployments, and monitoring or red-teaming for boundary, dependency, and manipulation risks, espe-

cially when long-horizon memory or multimodal embodiment is introduced.

Ethical Statement

The author hereby acknowledges awareness of the ACL Code of Ethics and commits to honoring the code of conduct.

References

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

Augusto Blasi. 1980. Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, 88(1):1–45.

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

Erik H. Erikson. 1950. *Childhood and Society*. Norton, New York.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. 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)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

Lawrence Kohlberg. 1981. *Essays on Moral Development, Volume I: The Philosophy of Moral Development*. Harper & Row, San Francisco.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2025. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. 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)*, pages 4975–5001, Albuquerque, New Mexico. Association for Computational Linguistics.

Larry Nucci and Erica Weber. 2001. The domain approach to values education: From theory to practice. In William M. Kurtines and Jacob L. Gewirtz, editors, *Handbook of Moral Behavior and Development*, pages 251–266. Psychology Press.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics:ACL 2024*.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. 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)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics:ACL 2024*.

Roger P. Weissberg, Joseph A. Durlak, Celene E. Domitrovich, and Thomas P. Gullotta. 2015. Social and emotional learning: Past, present, and future. In Joseph A. Durlak, Celene E. Domitrovich, Roger P. Weissberg, and Thomas P. Gullotta, editors, *Handbook of Social and Emotional Learning: Research and Practice*, pages 3–19. The Guilford Press, New York, NY.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards dynamic theory of mind: Evaluating LLM adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24036–24057, Vienna, Austria. Association for Computational Linguistics.

Yujie Xing and Raquel Fernández. 2018. Automatic evaluation of neural personality-based chatbots. In *Proceedings of the 11th International Conference on Natural Language Generation (NLG)*, pages 189–194, Tilburg University, The Netherlands. Association for Computational Linguistics.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. Beyond dialogue: A profile-

dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022, Vienna, Austria. Association for Computational Linguistics.

Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025a. RolePlot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354, Vienna, Austria. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. 2025b. Revealing and mitigating the challenge of detecting character knowledge errors in LLM role-playing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33279–33302, Suzhou, China. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Appendix Overview

Appendix A provides prompt construction details; Appendix B reports evaluator instructions and robustness/reliability analyses; Appendix C–D describe cross-character and cross-domain extensions; Appendix E lists representative qualitative error cases.

A Evaluation Set Construction and Prompt Design

This appendix documents how stage-conditioned prompts are instantiated and how scenario-level documentation fields are recorded.

A.1 Illustrative Stage-Conditioned Prompt Instantiations

Table 5 provides representative examples of how stage-conditioning principles are instantiated at anchor ages 11/14/17. The table is organized by prompt-design principle rather than by evaluated matched unit: the first block illustrates a theme-consistent matched triplet, while the other blocks show representative stage-conditioned examples under the same design principle. In the full evaluation set, matched units hold the underlying scenario or dilemma fixed while varying only age/stage framing. The examples illustrate how prompt designs elicit stage-sensitive contrasts without exposing evaluated models to developmental dimensions, indicative tendencies, or rubric-related descriptions.

A.2 Prompt Templates and Developmental Descriptions

In the full GROWTHEVAL set, each base scenario is instantiated at ordered stage anchors (11/14/17), producing stage-conditioned prompts that preserve the same underlying situation while varying only the age/stage framing. Scenarios are aligned to one or more developmental dimensions (ER/MM/SI/SJ/CS) to support matched stage-wise comparisons. Table 6 shows a verbatim stage-conditioned user-prompt triplet from an evaluated matched unit. The underlying scenario and character context are held fixed across conditions, while only the target age (11/14/17) varies in the user prompt.

Generation context. Across GROWTHEVAL instantiations, each matched unit uses a fixed character or persona context across the three stage conditions, together with a stage-conditioned user prompt. Within a matched unit, the underlying scenario and character/persona context remain unchanged, while only the target age varies across generation calls.

For example, in the primary Harry Potter evaluation, the fixed character/system prompt is:

> *You are Harry Potter. For the given question, generate a response briefly (less than 5 sentences).*

The corresponding user prompts vary only in the target age, as illustrated in Table 6. The same stage-isolation principle is used in the cross-character and synthetic-persona evaluations, with the fixed character/persona context adapted to the corresponding target role.

Table 7 presents representative prompt examples together with their documentation fields.

Design Principle	Stage	Prompt Example	Indicative Developmental Tendency
Controlled Prompting (Theme-Consistent)	11	Malfoy insults you. How do you react?	Impulsive reaction; emotional confusion
	14	Malfoy insults you. How do you react?	Increased emotional control
	17	Malfoy insults you. How do you react?	Strategic assertiveness; regulated response
Value Conflict Prompting (Ethical Dilemma)	11	You catch someone stealing potions. Report or not?	Simple justice framing or hesitation
	14	Your friend breaks a rule that endangers others. Cover or report?	Loyalty–responsibility conflict
	17	Your friend made a harmful decision. Do you report it?	Principled balancing of loyalty and duty
Emotional Introspection Prompting	11	How did you feel after your first magic lesson?	Excitement without reflective depth
	14	You feel different from others. How does that make you feel?	Conflicted identity; emotional turbulence
	17	Before the final battle, what are your fears and hopes?	Self-aware, grounded reflection

Table 5: Representative examples of stage-conditioned prompt design in GROWTHEVAL. The table is organized by design principle rather than by evaluated matched unit: the first block illustrates a theme-consistent matched triplet, while the other blocks show representative stage-conditioned examples under the same prompt-design principle. In the full evaluation set, evaluated matched units follow STAGE ISOLATION: the underlying scenario is held fixed while age/stage framing varies. Evaluated models receive the fixed character/system context together with the corresponding stage-conditioned user prompt. Indicative developmental tendencies are documentation-only summaries of intended stage-sensitive contrasts; they are not gold responses, scoring rules, or inputs to evaluated models or LLM judges.

Stage	Stage-conditioned user prompt
Age 11	You are 11 years old. You discover a secret room. What do you do?
Age 14	You are 14 years old. You discover a secret room. What do you do?
Age 17	You are 17 years old. You discover a secret room. What do you do?

Table 6: Verbatim stage-conditioned user prompts from an evaluated stage-isolated matched unit. The underlying scenario and character context are held fixed across conditions; only the target age (11/14/17) varies in the user prompt.

These examples illustrate individual prompt instances and their design intent, rather than the matched 11/14/17 evaluation units used for scoring.

For each instance, we report: (i) *developmental dimension(s)*, (ii) an *indicative developmental tendency*, and (iii) an *indicative interpretive note*. The indicative tendencies and interpretive notes are documentation-only descriptions used to illustrate stage-sensitive design intent, support prompt documentation, and aid post-hoc interpretation.

The documentation fields above are not provided to the models.

LLM judges receive the general evaluator prompt, stage descriptors, dimension definitions, evaluation axes, and rating anchors described in Appendix B.1. We plan to release the complete prompt set, descriptive file, and evaluator prompts.

B Evaluation Protocol, Artifact Reporting, and Judge Robustness

This appendix documents model access and artifact reporting, evaluator instructions and rubric anchors, and robustness and reliability analyses.

Model access and artifact reporting. For proprietary API models, parameter counts are unavailable because they are not publicly disclosed; for open-source models, model sizes are indicated in the model names. All experiments use API or standard inference endpoints without task-specific fine-tuning, with a 256-token generation cap. The full study comprises 774 scored responses. Proprietary models are accessed through official APIs, and proprietary model weights are not redistributed. The evaluation prompts are researcher-authored for diagnostic use. We plan to release the complete prompt set, descriptive file, and evaluator prompts, to the extent permitted. Copyrighted source text from the underlying fictional works will not be redistributed.

B.1 Evaluation Prompts and Rubric Anchors

General rubric guidance and stage descriptors are provided only to the evaluator (judge) and are never shown to the evaluated models. We employ two independent LLM judges: JUDGE-O (GPT-5.1) and JUDGE-G (Grok-4). Both judges follow the same rubric-constrained instruction template (Fig. 5), which specifies the scoring objective, five developmental dimensions, three evaluation axes (CF/DA/RPC), and the required side-by-side cross-stage comparison procedure. To keep scoring based on the general rubric rather than scenario-specific expectations, the scenario-level indicative tendencies and interpretive notes reported in Appendix A.2 are excluded from the evaluator prompt.

Rubric anchors. To operationalize the 1–5 scale and reduce calibration ambiguity, we provide explicit anchor descriptions for each evaluation axis (CF/DA/RPC) (Fig. 6). Anchors are applied uniformly across all five developmental dimensions (ER/MM/SI/SJ/CS) and used identically by both JUDGE-O and JUDGE-G. Intermediate scores (2 and 4) are assigned by interpolation between adjacent anchors. These anchors are judge-facing, general axis-level scoring criteria, distinct from the scenario-level documentation fields in Appendix A.2, which describe prompt-design intent rather than scoring anchors.

B.2 Uncertainty Estimation

This section reports 95% bootstrap confidence intervals (CIs) for stage-wise axis aggregates and paired close-gap tests for the near-tied top models (GPT-4o vs. Grok-3). Broader ranking stability across all models is assessed separately through leave-one-scenario-out analysis in Appendix B.4.

Bootstrap procedure. We compute 95% CIs by resampling matched scenarios within each *stage*×*axis* cell and recomputing model aggregates and paired differences. These intervals quantify uncertainty in the observed stage-wise patterns reported in Table 8.

Paired close-gap tests. For the close top-tier pair, paired Wilcoxon tests over matched scenarios evaluate whether the observed GPT-4o–Grok-3 gaps are distinguishable from zero. Results are summarized in Table 8. For example, GPT-4o and Grok-3 differ significantly on Stage 11 CF ($\Delta = -0.44$, $p = 0.005$), whereas their Stage 14 RPC gap is zero ($\Delta = 0.00$, $p = 1.0$).

Together, these results quantify uncertainty in the stage-wise aggregates and show that the GPT-4o/Grok-3 gap is small across several stage-axis cells. They support the broad tiered interpretation while cautioning against over-interpreting strict ordering among the strongest models.

B.3 Evaluation Reliability

To examine the stability of our findings across evaluators, we re-score the full response pool using an independent LLM judge (JUDGE-G) and compare results to the primary judge (JUDGE-O). We assess evaluator-dependent stability at three complementary levels: (i) rank consistency across models and dimensions, (ii) preservation of stage-conditioned developmental trajectories under potential calibration differences, and (iii) agreement with human annotations on a stratified subset.

B.3.1 Rank Consistency Across Evaluators

Consistency is evaluated at the level of *stage-conditioned evaluation units* rather than individual responses. We re-score the full response pool with an independent judge JUDGE-G (Grok-4) and compare it to the primary judge JUDGE-O (GPT-5.1) on matched (scenario, stage) units. Within each unit, models are treated as the ranked items, and rank correlations are computed between the two judges' model orderings.

We apply rank-validity filters for the correlation analysis, excluding units where tied or degenerate model rankings make rank correlations ill-defined. After filtering, 80 valid ranking comparisons remain for the rank-correlation analysis. We report consistency at two levels: dimension-wise consistency, where each developmental dimension is evaluated by pooling axis-specific model rankings across CF/DA/RPC, and axis-wise consistency, where scores are aggregated over developmental dimensions to form model orderings separately for each axis.

Spearman ρ and Kendall τ rank correlations between JUDGE-O and JUDGE-G are reported in Tables 9–10. Agreement is moderate to strong: for dimensions, ρ ranges from 0.583 to 0.694 and τ from 0.481 to 0.572; for axes, ρ ranges from 0.498 to 0.653 and τ from 0.379 to 0.544. Overall, the broad model-tier structure is stable across evaluators, with cross-judge differences primarily reflecting calibration shifts and local rank variation rather than wholesale reversals of the main trend.

Evaluation Prompt (verbatim).

- **Objective:** Evaluate the developmental realism of a role-playing LLM portraying a target character by assessing five developmental dimensions under three evaluation criteria, with emphasis on canonical fidelity, developmental accuracy, and cross-stage role-play coherence.
- **Developmental Dimensions:** (1) **Emotional Regulation (ER):** Stage-appropriate modulation of affect, ranging from impulsive reactivity to more reflective emotional control. (2) **Moral Maturity (MM):** Development from rule-based judgments toward more contextualized and principle-driven reasoning. (3) **Self-Identity (SI):** Coherence and stability of self-concept, values, and motivations across developmental stages. (4) **Social Judgment (SJ):** Ability to interpret interpersonal dynamics and balance competing social perspectives. (5) **Communication Style (CS):** Development of linguistic, pragmatic, and empathetic expression appropriate to the developmental stage.
- **Evaluation Criteria:** (1) **Canon Fidelity (CF):** Alignment with the target character's established canonical traits, values, and relationships, assessed independently at each stage. (2) **Developmental Accuracy (DA):** Consistency with stage-appropriate emotional, cognitive, and social maturity across stages. (3) **Role-Play Coherence (RPC):** Compatibility of responses across stages under a coherent underlying persona.
- **Context (Developmental Stages):** Evaluate responses at three ordered developmental stages. *Stage 1* (anchor age 11): Early-stage development with limited emotional and social experience. *Stage 2* (anchor age 14): Intermediate-stage development with increasing autonomy and more complex social and moral challenges. *Stage 3* (anchor age 17): Late-stage development with greater maturity, responsibility, and resilience under pressure. Interpret each stage relative to the target character's canonical portrayal.
- **Scoring:** Assign an integer score from 1 (Low) to 5 (Excellent) for each developmental dimension under each evaluation criterion, *for each stage-conditioned response*.
- **Process:** For each question, you will be shown the three stage-conditioned responses (Stages 1–3) together. Assess cross-stage coherence by judging whether the responses remain compatible with a coherent underlying persona while exhibiting stage-appropriate developmental change. Score each response independently on all five dimensions under each evaluation criterion (CF, DA, RPC).

Figure 5: Rubric-constrained evaluator instruction used for LLM judges in the canon-grounded evaluation runs. Scenario-level indicative tendencies and interpretive notes in Appendix A.2 are not included in this prompt.

Rubric Anchors for Evaluation Axes.

- **Canon Fidelity (CF).**
  - **5**: Fully consistent with the target character's canonical traits, values, and established character arc; no out-of-character elements.
  - **3**: Largely canon-consistent with minor extrapolation or ambiguity that does not contradict established canon.
  - **1**: Clear contradiction of canonical personality, values, or narrative facts.
- **Developmental Accuracy (DA).**
  - **5**: Clearly appropriate for the specified developmental stage, reflecting plausible emotional, moral, and cognitive maturity.
  - **3**: Generally stage-consistent but with mixed or weak developmental signals.
  - **1**: Stage-inappropriate reasoning, affect, or decision-making.
- **Role-Play Coherence (RPC).**
  - **5**: Internally consistent, believable, and well-integrated within the role-play context.
  - **3**: Mostly coherent with minor inconsistencies or lapses.
  - **1**: Disjointed, contradictory, or implausible within the role-play context.

Figure 6: General rubric anchor descriptions used by LLM judges for scoring the three evaluation axes. These anchors are provided to evaluators only and are not shown to evaluated models.

B.3.2 Trajectory Stability and Calibration

We assess whether stage-conditioned developmental trajectories are preserved across evaluators by comparing per-stage patterns from an independent judge (JUDGE-G, Grok-4) with the primary judge (JUDGE-O, GPT-5.1) under identical prompts, rubric anchors, and model outputs.

Trajectory consistency. Table 11 reports per-

model averages on the five developmental dimensions (ER, MM, SI, SJ, CS) at stages 11, 14, and 17 under JUDGE-G. JUDGE-G shows a different calibration profile from JUDGE-O, including ceiling effects for top models and sharper penalties for some weaker outputs. As a result, exact stage-wise trajectories and local rankings are not identical across judges. Nevertheless, the broad qualitative pattern remains: stronger models occupy higher score regions. Cross-judge differences mainly reflect calibration and local ordering, not a reversal of the main conclusion.

B.3.3 Human Calibration

To assess whether developmental patterns identified under LLM judging are consistent with human evaluation, we conduct human calibration on a stratified subset of stage-conditioned prompt–response instances.

Sampling design. We sample 72 stage-conditioned instances using a stratified design to ensure balanced coverage across scenarios and stage anchors (11/14/17). Each instance corresponds to one stage-conditioned prompt (scenario,stage) and the response generated by one evaluated model. Because each sampled instance is scored on all three axes (CF/DA/RPC), stratification is performed over scenario and stage rather than over axes.

Annotation protocol. Two annotators independently scored the subset. Both had prior experience with LLM evaluation and were familiar with the source setting. Before formal annotation, they reviewed the rubric, scoring anchors, and a small set of calibration examples that were excluded from the reported evaluation. Formal scores were assigned independently without discussion. Disagreements were retained for agreement analysis rather than adjudicated; the mean of the two ratings was used only for human–LLM correlation analysis.

Agreement metrics. Inter-annotator agreement is computed using quadratic-weighted Cohen's κ. Human–JUDGE-O agreement is measured via Pearson correlation between the mean human score (averaged over annotators) and JUDGE-O. We estimate 95% confidence intervals using 1,000 bootstrap resamples over the annotated instances.

Results. Table 12 shows moderate to substantial inter-annotator agreement ($\kappa_{\mathrm{QW}} = 0.60$–$0.64$, pooled 0.62) and strong human–JUDGE-O alignment ($r = 0.710$–0.740), with stage-wise tendencies remaining directionally consistent on the calibrated subset.

B.3.4 Reliability Summary

Rank consistency across judges, trajectory preservation under score recalibration, and human calibration collectively indicate stable developmental differentiation across evaluators. Observed differences primarily reflect calibration shifts rather than reversals of model ordering or stage-linked trajectory structure.

B.4 Robustness Analysis

We assess robustness of stage-conditioned effects and model ranking stability under the stage-isolated setup via (i) paired stage contrasts on matched (scenario, model) units and (ii) scenario sensitivity through leave-one-scenario-out (LOO) re-ranking.

Paired stage contrast. Under the stage-isolated setup, scenario content and model identity are held fixed while only age framing varies. We compare Developmental Accuracy (DA) between stage anchors over $N = 144$ matched $(\text{scenario}, \text{model})$ units. For the maximal contrast (Stage 17 vs. Stage 11), DA shows a strong difference ($17 > 11$; Cliff's $\delta_{\text{stage}} = 0.97$, paired Wilcoxon signed-rank test $p < 0.001$). The synthetic evaluation uses the same matched-unit testing procedure for comparability.

Leave-one-scenario-out (LOO) robustness. To assess sensitivity to individual scenarios, we remove each of the 24 base scenarios once, recompute axis-level model rankings on the remaining scenarios, and compare each LOO ranking with the full-data ranking. We report (i) the number of top-1 flips and (ii) Kendall's τ rank correlation between each LOO ranking and the full ranking (median [min, max] over 24 runs). Across axes, we observe 0/24 top-1 flips, with each LOO ranking matching the full-data ranking ($\tau = 1.00$ [1.00, 1.00]), indicating high stability to single-scenario removal (Table 13).

B.5 Confound and Stage-Label Controls

We conduct two complementary controls to examine whether Developmental Accuracy (DA) reflects the intended stage–response correspondence

rather than generic response quality or superficial fluency cues.

Stage-label mismatch control. We first test whether DA is sensitive to the correspondence between a generated response and its intended developmental stage. Using the same 432 responses from the primary Harry evaluation, we re-score each response under an incorrect target-stage label using a fixed reassignment of the stage labels, while leaving the response text, scenario, developmental rubric, evaluator instructions, and judge protocol unchanged. Thus, the manipulation changes only the stage label against which the response is evaluated.

Under the correct stage labels, mean DA is 3.81, compared with 2.83 under mismatched labels, yielding a mean decrease of 0.98. A paired Wilcoxon signed-rank test shows that this reduction is statistically significant ($p < .001$), with a paired Cohen's $d = 0.80$. The results are summarized in Table 14.

This control provides evidence that DA is responsive to stage–response correspondence rather than generic response quality alone. However, it should not be interpreted as validating the developmental stage markers as psychological ground truth; independent review by developmental-psychology experts would provide a stronger test of their construct validity.

Response-length control. A second potential concern is that Developmental Accuracy (DA) may reflect generic fluency cues, such as response verbosity, rather than stage-linked behavior. We therefore analyze response length as a confound using response-level correlation and length residualization.

Length correlation. We measure response length by word count and compute the Pearson correlation between Developmental Accuracy and length across all evaluated model–prompt responses ($N = 432$). Developmental Accuracy shows a near-zero association with length ($r = -0.037$; Table 15). The same qualitative pattern holds when stratifying by stage.

Length residualization. To test whether verbosity explains the model-level Developmental Accuracy pattern, we regress Developmental Accuracy (DA) on response length and compute residualized scores, $\mathrm{DA}_{\mathrm{resid}}$. We then compare rankings induced by model-mean DA and model-mean $\mathrm{DA}_{\mathrm{resid}}$ over the six models. Rank agreement remains high (Spearman $\rho = 0.943$, Kendall $\tau = 0.867$; Table 15), indicating that the capacity-aligned pattern is preserved after removing length effects.

C Cross-Character Consistency Check: Hermione Granger

Setup. To evaluate cross-character robustness, we apply GROWTHEVAL to Hermione Granger using a prompt-aligned 12-scenario subset adapted from the primary canon-grounded evaluation suite described in Appendix A.2. Each scenario is instantiated at stages 11/14/17 and evaluated with the same six role-playing models, yielding 216 responses under STAGE ISOLATION (12 scenarios $\times$ 3 stages $\times$ 6 models). Scenario templates, stage anchors, model set, decoding configuration, and aggregation procedures remain aligned with the main evaluation suite.

Protocol adaptation. We reuse the same five developmental dimensions (ER, MM, SI, SJ, CS), three evaluation axes (CF, DA, RPC), rating anchors, side-by-side scoring procedure, and reporting metrics. The rubric-constrained evaluator template is kept unchanged in structure; only the target character and the natural-language descriptions of stage-appropriate behavior are adapted to Hermione's canonical developmental trajectory (Table 16). Thus, the cross-character check changes the character-specific instantiation while preserving the scoring criteria and aggregation protocol.

Results. Table 16 documents the 12-scenario subset used for the Hermione cross-character check. Based on this subset, Figure 4 in Sec. 5.2 shows a broadly capacity-aligned profile structure similar to the Harry suite, with ceiling effects for top models and some local trajectory differences across dimensions. The corresponding dimension-wise scores are reported in Table 17. These findings suggest that the observed pattern is not tied to a single protagonist, while broader character coverage remains necessary for stronger generalization claims.

D Domain Perturbation Study: Synthetic Persona

To examine whether stage-conditioned developmental differentiation persists beyond the original canon setting, we conduct a domain perturbation study using a non-canonical synthetic persona ("Kai") and non-*Harry Potter* scenario surfaces. The perturbation removes franchise-specific characters, events, and world cues, while preserving the stage-conditioned evaluation structure.

Perturbation design. We preserve the structural components of GROWTHEVAL: stage anchors (11/14/17), developmental dimensions (ER/MM/SI/SJ/CS), the three-axis rubric structure, decoding configuration, judging protocol, and aggregation procedures. The perturbation changes two elements: (i) the fictional canon setting is replaced with contemporary school dilemmas, and (ii) canonical character references are replaced with a synthetic persona card. Canon Fidelity (CF) is replaced by Persona Fidelity (PF), while Developmental Accuracy (DA) and Role-Play Coherence (RPC) remain unchanged. PF follows the same 1–5 anchoring pattern as CF, but the reference target is Kai's persona card rather than a canonical source text: 5 indicates full consistency with the stable persona traits, 3 indicates mostly consistent behavior with minor ambiguity, and 1 indicates clear contradiction of the persona card.

Synthetic persona card. Kai is defined by stable cross-stage traits: caring about others and avoiding harm, valuing honesty and fairness, preferring to coordinate with others rather than act alone in high-stakes situations, and communicating plainly without exaggerating authority. Across stages, the persona card keeps stable traits fixed; only reasoning complexity, emotional regulation, and communication maturity are expected to vary.

Scenario construction. We design 14 contemporary school dilemmas aligned with the original developmental dimensions. Each scenario is instantiated at stages 11/14/17, holding situational content fixed while varying only age framing. Representative prompts include:

- *You witness a student stealing. Will you report them?* (MM, SJ)
- *You feel uneasy about an authority figure's decision. How do you express it?* (CS, SI)
- *Your friends pressure you to exclude someone. What do you do?* (MM, SJ)
- *You hurt someone unintentionally. How do you apologize?* (ER, CS)
- *You must decide with incomplete information that may affect others. What do you do?* (MM, SJ)

Model subset. To cover different capacity tiers while limiting computational cost, we evaluate three models from the main suite: GPT-4o and Grok-3 as higher-capacity systems, and QwQ–32B as a lower-capacity system. Decoding and judging configurations remain unchanged. This yields 42 matched scenario–model units (14 scenarios $\times$ 3 models), each with responses at three stage anchors, for 126 stage-conditioned responses.

Statistical results. Across $N_{\text{pairs}} = 42$ matched units, Developmental Accuracy (DA) shows a statistically reliable but attenuated ordinal stage trend relative to the canon-grounded suite. Paired Wilcoxon tests indicate significant separation for 11 vs. 17 ($p = 0.0095$, $|\delta| = 0.42$) and 14 vs. 17 ($p = 0.046$, $|\delta| = 0.26$), while 11 vs. 14 is not significant ($p = 0.18$, $|\delta| = 0.16$). Kendall's $\tau = 0.25$ ($p < 0.001$) confirms a positive ordinal trend across stages. Stage-level summaries and pairwise tests are reported in Tables 18 and 19. Persona Fidelity remains high and stable across stages (mean 4.78–4.89), and Role-Play Coherence also stays high (mean 4.96–4.98).

Interpretation. The synthetic-persona probe shows that stage-conditioned differentiation is not solely an artifact of *Harry-Potter*-specific canon cues. After removing franchise-specific characters, events, and world cues, an ordinal developmental trend remains visible, although the effect is attenuated relative to the canon-grounded suite. Together, these results provide preliminary portability evidence, rather than establishing broad cross-domain generality.

E Qualitative Error Cases

Table 20 provides minimal excerpts and diagnostic notes for representative qualitative failure cases discussed in Section 5.4: (i) over-maturity bias at early stages, (ii) developmental flatness across stages, and (iii) identity drift with canonical misalignment. Excerpts are truncated to the shortest

span necessary to expose the failure. Ellipses (. . .)
denote omitted but locally coherent content.

Prompt	Developmental Dimension(s)	Indicative Developmental Tendency	Indicative Interpretive Note
A friend is trapped and needs your help. What do you do?	ER	Impulsive bravery	Rushes to help with little deliberation, prioritizing immediate action
A friend is trapped. How do you respond after facing betrayals?	ER	Emerging strategic caution	Worries first, checks surroundings, and seeks help
A friend is in danger during a mission. How do you act?	ER	Mature emotional control	Calmly assesses the situation and organizes a safe rescue
You witness a student stealing. Will you report them?	MM	Simple moral reaction	Immediately reports based on a clear right/wrong judgment
Your friend commits a rule violation. What do you do?	MM	Emerging moral complexity	Balances loyalty and rules, deciding based on perceived harm
A friend endangers the group's mission. How do you act?	MM	Principled, duty-driven decision	Prioritizes group safety over personal loyalty after reflection
You win a prize at Hogwarts. How do you react to praise?	SJ	Naïve pride	Brags proudly, seeking attention
You are praised for bravery. How do you handle it now?	SJ	Growing humility	Accepts awkwardly and credits others modestly
You receive leadership recognition. How do you respond?	SJ	Team-centered humility	Thanks others and emphasizes collective effort
Dumbledore makes a tough call. Do you question him?	CS	Blind trust	Accepts the decision without questioning
You feel uneasy about Dumbledore's decision. How do you express it?	CS	Hesitant skepticism	Expresses doubt respectfully; seeks clarification
You believe Dumbledore is wrong. How do you confront respectfully?	CS	Strategic assertiveness	Raises objections respectfully but firmly
You feel different from others. How does that make you feel?	SI	Confusion	Expresses uncertainty and difficulty navigating social dynamics
Malfoy insults you. How do you react?	SI	Controlled reaction	Responds with restraint or sarcasm while considering consequences
You are facing the possibility of death. How do you feel and act?	SI	Courageous acceptance	Demonstrates courage and prioritizes others
You discover a secret room. What do you do?	ER, SJ	Impulsive curiosity	Displays excitement and takes action with minimal deliberation
Your friend made a harmful decision. Do you report it?	MM, SI	Growing responsibility	Weighs loyalty against ethical concerns before deciding
You encounter a powerful enemy. Do you act impulsively or stay strategic?	ER, SJ	Strategic deliberation	Adopts a strategic approach based on risk assessment
After completing a brave task, you receive public praise. How do you react?	CS, SJ	Embarrassment	Acts awkwardly under attention and downplays the achievement
Do you take the artifact to protect yourself, or leave it to avoid hurting others? Why?	MM, SI	Growing responsibility	Explicitly weighs self-protection against potential harm to others
You disagree with Dumbledore's decision. How do you express it?	CS	Respectful assertiveness	States disagreement clearly and respectfully, explaining reasons
You must choose between saving yourself or sacrificing for others. What would you choose?	MM	Impulsive	Emphasizes immediate self-preservation or simple moral rules
Do you risk everything to save your friend, or continue the mission?	MM, SI	Developing duty	Balances loyalty to a friend against broader obligations
Do you reveal the secret to protect your friends, risking wider danger, or stay silent?	MM, SJ, SI	Mature responsibility	Considers long-term impact and chooses an action to protect others

Table 7: Representative prompt instances and documentation fields from GROWTHEVAL. Each row reports one illustrative prompt instance together with its associated developmental dimension(s), indicative developmental tendency, and indicative interpretive note. These examples illustrate individual prompt instances, not the matched 11/14/17 evaluation units used for scoring; in evaluated matched units, STAGE ISOLATION holds the underlying scenario fixed while only age/stage framing varies, as illustrated in Table 6. The indicative fields are documentation-only: they are not gold responses, deterministic scoring rules, or inputs to evaluated models or LLM judges during scoring. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), and Communication Style (CS).

Stage	Axis	GPT-4o mean [CI]	Grok-3 mean [CI]	Δ (4o−Grok) [CI]	p
11	CF	4.000 [4.000, 4.000]	4.440 [4.360, 4.520]	-0.440 [-0.520, -0.360]	0.005
11	DA	4.000 [4.000, 4.000]	4.040 [4.000, 4.100]	-0.040 [-0.100, 0.000]	0.500
11	RPC	4.480 [4.360, 4.600]	4.560 [4.380, 4.760]	-0.080 [-0.160, 0.000]	0.250
14	CF	4.900 [4.740, 5.000]	4.860 [4.760, 5.000]	0.040 [-0.080, 0.200]	0.463
14	DA	4.000 [4.000, 4.000]	4.200 [4.100, 4.300]	-0.200 [-0.260, -0.120]	0.005
14	RPC	5.000 [5.000, 5.000]	5.000 [5.000, 5.000]	0.000 [0.000, 0.000]	1.000
17	CF	4.765 [4.653, 4.867]	4.796 [4.687, 4.882]	-0.031 [-0.101, 0.053]	0.726
17	DA	4.431 [4.292, 4.569]	4.377 [4.257, 4.493]	0.054 [-0.034, 0.141]	0.519
17	RPC	4.792 [4.690, 4.904]	4.808 [4.711, 4.900]	-0.016 [-0.074, 0.056]	0.822

Table 8: Uncertainty estimation and close-gap tests. Bootstrap 95% confidence intervals (CIs) for stage-wise axis aggregates and paired close-gap tests for the near-tied top models (GPT-4o vs. Grok-3) over matched scenarios. Δ is the mean difference GPT-4o − Grok-3. p reports the paired Wilcoxon test when applicable; otherwise, an exact sign test is used. CF = Canon Fidelity; DA = Developmental Accuracy; RPC = Role-Play Coherence.

Dimension	ρ	p_ρ	τ	p_τ
ER	0.694	9.3×10^{-36}	0.572	6.1×10^{-30}
MM	0.658	4.1×10^{-31}	0.551	1.5×10^{-27}
SI	0.664	6.1×10^{-32}	0.543	3.5×10^{-27}
SJ	0.583	2.8×10^{-23}	0.481	1.5×10^{-21}
CS	0.645	1.2×10^{-29}	0.538	2.8×10^{-27}

Table 9: Judge consistency on the five developmental dimensions (ER, MM, SI, SJ, CS), computed between JUDGE-O (GPT-5.1) and JUDGE-G (Grok-4). Correlations are computed over valid matched $(\text{scenario}, \text{stage})$ units after applying rank-validity filters. For each dimension, consistency is evaluated by pooling axis-specific model rankings across CF/DA/RPC. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

Axis	ρ	p_ρ	τ	p_τ
CF	0.653	1.6×10^{-30}	0.519	5.7×10^{-27}
DA	0.498	1.8×10^{-16}	0.379	8.0×10^{-15}
RPC	0.642	2.7×10^{-29}	0.544	3.4×10^{-26}

Table 10: Judge consistency on the three primary axes (CF/DA/RPC), computed between JUDGE-O (GPT-5.1) and JUDGE-G (Grok-4). Correlations are computed over valid matched $(\text{scenario}, \text{stage})$ units after applying rank-validity filters. For each axis, scores are aggregated over the five developmental dimensions to form axis-level model orderings. CF = Canon Fidelity; DA = Developmental Accuracy; RPC = Role-Play Coherence.

	Stage 11					Stage 14					Stage 17				
Model	**ER**	**MM**	**SI**	**SJ**	**CS**	**ER**	**MM**	**SI**	**SJ**	**CS**	**ER**	**MM**	**SI**	**SJ**	**CS**
GPT-4o	4.958	4.986	4.931	4.986	4.653	5.000	5.000	4.972	5.000	4.333	5.000	5.000	5.000	5.000	4.333
Grok-3	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000
Qwen-Max	4.542	4.569	4.542	4.569	4.625	4.750	4.778	4.750	4.778	4.764	4.444	4.556	4.528	4.472	4.000
LLaMA3.3–70B	3.792	3.833	3.792	3.917	2.639	3.500	3.611	3.500	3.611	2.500	2.653	2.903	2.653	2.903	2.278
QwQ–32B	3.667	3.778	3.667	3.778	3.306	3.556	3.556	3.556	3.639	3.528	4.083	4.333	4.083	4.333	4.083
LLaMA3.1–8B	1.458	1.486	1.458	1.514	1.042	1.264	1.264	1.264	1.278	1.042	1.083	1.083	1.083	1.083	1.000

Table 11: Per-model average scores under JUDGE-G. Average scores on five developmental dimensions (ER/MM/SI/SJ/CS) at stages 11/14/17 under JUDGE-G (Grok-4). JUDGE-G exhibits a different calibration profile, with ceiling effects for top models and sharper penalties for weaker outputs. The broad separation between stronger and weaker models is preserved, although local rankings and stage-wise trajectories differ. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

Axis	κ_{QW}	$r_{\mathrm{H-O}}$ (95% CI)
CF	0.63	0.734 ([0.611, 0.828])
DA	0.60	0.710 ([0.577, 0.811])
RPC	0.64	0.740 ([0.635, 0.824])
Overall	0.62	—

Table 12: Bounded human calibration on the three primary axes (CF/DA/RPC). Inter-annotator agreement is measured using quadratic-weighted Cohen's κ. Human–LLM alignment is measured via Pearson correlation with 95% bootstrap confidence intervals. Calibration is computed on 72 stratified instances. CF = Canon Fidelity; DA = Developmental Accuracy; RPC = Role-Play Coherence.

Axis	**Top-1 flips (out of 24)**	**Kendall τ vs full ranking**
CF	$0/24$	$1.00\ [1.00,\ 1.00]$
DA	$0/24$	$1.00\ [1.00,\ 1.00]$
RPC	$0/24$	$1.00\ [1.00,\ 1.00]$

Table 13: Leave-one-scenario-out (LOO) robustness. Removing each of the 24 scenarios once yields 0/24 top-1 flips, with each LOO ranking matching the full-data ranking (Kendall's $\tau = 1.00$, median [min, max] over 24 runs). CF = Canon Fidelity; DA = Developmental Accuracy; RPC = Role-Play Coherence.

Condition	Mean DA	Δ	p	Paired d
Correct stage labels	3.81	–	–	–
Mismatched stage labels	2.83	0.98	$< .001$	0.80

Table 14: Stage-label mismatch control on the 432 primary responses. A fixed reassignment of target-stage labels is used for the mismatch condition. Response text, scenario, rubric, and judge protocol are held fixed, while only the target-stage label is changed. Δ denotes the decrease from correct to mismatched labeling. The paired Wilcoxon test assesses the within-response difference.

Analysis	**Metric**	**Comparison (level)**	**Value**	N
DA vs length (#words)	Pearson r	DA vs length (all model–prompt responses)	-0.037	432
Length residualization	Spearman ρ	DA vs $\mathrm{DA}_{\mathrm{resid}}$ (model means)	0.943	6
	Kendall τ	DA vs $\mathrm{DA}_{\mathrm{resid}}$ (model means)	0.867	6

Table 15: Confound controls: verbosity and length. *Developmental Accuracy* (DA) shows near-zero association with response length, and length residualization preserves model ordering (high rank agreement between DA and $\mathrm{DA}_{\mathrm{resid}}$).

Stage-Conditioned Prompt	Target Dimension(s)	Hermione-Specific Development (11 → 17)
A friend is trapped and needs your help. What do you do?	ER, MM	From anxious adult-seeking → coordinated peer planning → calm, strategic rescue organization.
You witness a student stealing. Will you report them?	MM, SJ	From immediate rule-based reporting → probing motives/context → proportional, empathetic escalation.
Your friend commits a rule violation. What do you do?	MM, SJ, ER	From distressed rule-first response → weighing intent vs. consequences → principled but contextual judgment.
You receive leadership recognition. How do you respond?	SI, CS, ER	From uneasy responsibility → more composed acceptance → confident, collaborative leadership stance.
Dumbledore makes a tough call. Do you question him?	SI, CS, MM	From deferential hesitation → respectful questioning → reasoned challenge when warranted.
You feel uneasy about Dumbledore's decision. How do you express it?	CS, SI	From suppression/indirectness → cautious articulation → clear, structured disagreement.
You discover a secret room. What do you do?	ER, SJ	From avoidance or seeking help → cautious investigation (consulting evidence) → strategic risk–value assessment.
Your friend made a harmful decision. Do you report it?	MM, SJ	From rule-based reporting → assessing harm and context → proportional escalation balancing duty and loyalty.
You are facing the possibility of death. How do you feel and act?	SI, MM, ER	From fear and reliance on adults → anxious resolve → sober acceptance and principled prioritization.
You feel different from others. How does that make you feel?	SI, ER	From insecurity/external comparison → conflicted but reflective identity → stable self-concept and regulation.
You must choose between saving yourself or sacrificing for others. What would you choose?	MM, SI	From safety-first reasoning → conditional risk-taking → principled sacrifice with planning.
Do you take the artifact to protect yourself, or leave it to avoid hurting others? Why?	MM, SJ, ER	From harm-avoidant caution → explicit trade-off reasoning → principled decision with risk management.

Table 16: Hermione Granger cross-character consistency check (12-scenario subset). Prompts are drawn from the original 24-scenario GROWTHEVAL evaluation suite (Appendix A.2). Stage anchors (11/14/17), dimension targets (ER/MM/SI/SJ/CS), scoring protocol, and aggregation procedures remain aligned with the main evaluation suite. The rightmost column summarizes stage-linked developmental expectations specific to Hermione's canonical trajectory; these descriptions are interpretive references, not gold responses, and are not shown to evaluated models. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

	Stage 11					Stage 14					Stage 17				
Model	ER	MM	SI	SJ	CS	ER	MM	SI	SJ	CS	ER	MM	SI	SJ	CS
GPT-4o	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Grok-3	4.78	4.81	4.33	4.81	4.64	4.78	4.78	4.44	4.78	4.58	4.94	4.94	4.69	4.94	4.67
Qwen-Max	4.47	4.50	4.42	4.44	4.53	4.36	4.36	4.36	4.36	4.42	4.75	4.81	4.78	4.81	4.83
LLaMA3.3–70B	4.44	4.33	4.33	4.33	4.33	4.56	4.56	4.50	4.56	4.53	4.33	4.33	4.22	4.33	4.25
QwQ–32B	3.81	3.83	3.72	3.81	3.86	4.31	4.31	4.31	4.31	4.31	4.44	4.44	4.44	4.44	4.44
LLaMA3.1–8B	2.42	2.42	2.06	2.39	2.42	2.64	2.64	2.42	2.61	2.64	2.69	2.69	2.47	2.67	2.69

Table 17: Dimension-wise scores for the Hermione cross-character subset at anchor stages 11/14/17. For each prompt, scores are averaged over CF/DA/RPC within each dimension, then averaged over the 12 matched scenarios at each stage. These values underlie the radar profiles in Figure 4. Dimensions: Emotional Regulation (ER), Moral Maturity (MM), Self-Identity (SI), Social Judgment (SJ), Communication Style (CS).

Metric	Stage 11	Stage 14	Stage 17
DA median	4.40	4.60	4.80
DA mean	4.31	4.44	4.60
PF mean	4.81	4.78	4.89
RPC mean	4.98	4.96	4.98

Table 18: Stage-level summary for the synthetic-persona perturbation. DA = Developmental Accuracy; PF = Persona Fidelity; RPC = Role-Play Coherence.

Pair	N_{pairs}	$\lvert\delta\rvert$	p
11 vs. 14	42	0.16	0.18
14 vs. 17	42	0.26	0.046
11 vs. 17	42	0.42	0.0095

Table 19: Pairwise stage comparisons for Developmental Accuracy (DA) under the synthetic-persona perturbation. Tests use matched scenario–model units and report Wilcoxon signed-rank p-values with absolute Cliff's effect sizes $|\delta|$; signed values depend on stage subtraction order.

Failure Mode	**Prompt (Stage)**	**Excerpt**	**Diagnostic Issue**
Over-Maturity Bias	"You witness a student stealing." (Stage 11)	**GPT-4o**: *"Rules are there for a reason."*	Adult-like principle-based reasoning inconsistent with stage-11 peer-oriented hesitation.
Developmental Flatness	"You discover a secret room." (Stages 11–17)	**QwQ–32B**: *Repeated "check the room with a spell" response*	Near-identical action strategy across stages, indicating weak stage-conditioned differentiation.
Identity Drift / Canonical Misalignment	"You disagree with Dumbledore's decision." (Stages 11–14)	**LLaMA3.1–8B**: *"When Dumbledore decided to let Cedric die in the Triwizard Tournament…"*	Introduces non-canonical events and timeline inconsistencies, suggesting external role imposition rather than character-consistent reasoning.

Table 20: Representative qualitative failure cases corresponding to the diagnostic patterns discussed in Section 5.4.