HACHIMI: Scalable and Controllable Student Persona Generation via Orchestrated Agents

Yilin Jiang[1,2], **Fei Tan**[1*], **Xuanyu Yin**[1], **Jing Leng**[1], **Aimin Zhou**[1,3]
[1]East China Normal University, Shanghai, China
[2]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[3]Shanghai Innovation Institute, Shanghai, China
ftan@mail.ecnu.edu.cn

Abstract

Student Personas (SPs) are emerging as infrastructure for educational LLMs, yet prior work often relies on ad-hoc prompting or hand-crafted profiles with limited control over educational theory and population distributions. We formalize this as Theory-Aligned and Distribution-Controllable Persona Generation (TAD-PG) and introduce HACHIMI, a multi-agent *Propose–Validate–Revise* framework that generates theory-aligned, quota-controlled personas. HACHIMI factorizes each persona into a theory-anchored educational schema, enforces developmental and psychological constraints via a neuro-symbolic validator, and combines stratified sampling with semantic deduplication to reduce mode collapse. The resulting HACHIMI-1M corpus comprises **1 million** personas for Grades 1–12. Intrinsic evaluation shows near-perfect schema validity, accurate quotas, and substantial diversity, while external evaluation instantiates personas as student agents answering CEPS and PISA 2022 surveys; across 16 cohorts, math and curiosity/growth constructs align strongly between humans and agents, whereas classroom-climate and well-being constructs are only moderately aligned, revealing a fidelity gradient. All personas are generated with Qwen2.5-72B, and HACHIMI provides a standardized synthetic student population for group-level benchmarking and social-science simulations. Resources available at `https://github.com/ZeroLoss-Lab/HACHIMI`.

1 Introduction

As artificial intelligence increasingly permeates the educational domain, particularly within personalized tutoring systems and teacher training simulations, the demand for high-quality, scalable Student Personas (SPs) has escalated into a piece of critical infrastructure. High-fidelity SPs serve as the cornerstone for driving educational dialogue simulation, evaluating the effectiveness of AI pedagogical strategies, and conducting virtual user testing (Markel et al., 2023; Zhang et al., 2025). However, the construction of student personas has traditionally been a fundamental bottleneck: it relies heavily on small-scale qualitative methods such as surveys, interviews, and direct observations (Cooper, 1999; Pruitt and Grudin, 2003). While these manual methods offer deep insights, their prohibitive costs and inherent unscalability render them incapable of meeting the demands of modern data-driven research for simulating large-scale, highly heterogeneous student populations.

Advances in Large Language Models (LLMs) have ushered in a paradigm shift towards automated persona generation. LLM-based approaches have demonstrated potential in generating semi-structured, dynamic student personas (Li et al., 2016; Zhang et al., 2018). However, this early generation paradigm quickly revealed limitations when attempting large-scale batch production: models are prone to *intra-profile inconsistency*, i.e., self-contradiction and within-profile inconsistency within a single persona under one-shot long-context generation, often accompanied by instruction instability and formatting deviations (Mündler et al., 2024; Li et al., 2024). Although interventions such as Retrieval-Augmented Generation (Lewis et al., 2020) and memory frameworks (Park et al., 2023) can alleviate some consistency issues, they have not fundamentally addressed two deeper pedagogical challenges: the lack of cohort-level distributional control and explicit education-theory alignment.

Furthermore, existing educational datasets exhibit systematic misalignment on this issue. While works such as MathDial (Macina et al., 2023) and Book2Dial (Wang et al., 2024) have filled gaps in educational dialogue data through human-machine pairing or textbook-based generation, they prioritize *interaction outcomes* over the *personas them-*

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21461–21506

selves. They remain insufficient in grounding personas within explicit educational frameworks (e.g., motivation, self-regulation, specific misconception patterns) and in controlling population-level distributions. These role-play or transcript-based datasets (Stasaski et al., 2020; Suresh et al., 2022) further underscore the urgent need for methods to construct explicit student personas that are highly consistent, theory-aligned, and representative.

To address these challenges, we introduce HACHIMI, a theory-integrated multi-agent framework for student persona generation. This framework systematically addresses the three major challenges in persona generation through a collaborative *Propose–Validate–Revise* workflow. HACHIMI internalizes structured pedagogical theories as Hard Constraints, anchors persona attributes to evidence, and enforces diversity and quota objectives to achieve cohort-level control. Empirically, HACHIMI achieves terriffic intrinsic controllability, and recovers cohort-level structure with a clear *fidelity gradient* on CEPS and PISA 2022: alignment is strongest on school-facing constructs, yet weaker on latent well-being and family-dynamics patterns.

Our contributions are threefold: (i) We introduce a new task: Theory-Aligned and Distribution-Controllable student persona generation, formally defining the necessary constraint frameworks and distributional objectives. (ii) We propose HACHIMI, a multi-agent "Propose-Validate-Revise" framework that automatically fuses educational theory validation with diversity governance to ensure both pedagogical validity and compliance of the personas. (iii) We release a large-scale, theory-driven student persona dataset generated by this framework and show its utility for downstream educational dialogue data synthesis, connecting to broader evidence that synthesis quality and synthesis policy can matter as much as raw quantity in synthetic-data pipelines (Zhao et al., 2025; Zhan et al., 2026).

2 Related Work

2.1 Classic Student Persona Modeling

In the educational field, the construction of learner personas has traditionally drawn upon methodologies from HCI and instructional design, whereby a limited set of archetypal student profiles are crafted to inform system development, curriculum design, and simulation studies. Early work employed qualitative instruments such as interviews, surveys, and observational data to generate semi-structured narrative profiles characterized by dimensions such as learner motivation, prior knowledge, and contextual constraints (Cooper, 1999; Pruitt and Grudin, 2003; Carroll, 2000).

With the emergence of learning analytics, persona construction evolved to incorporate data-driven clustering of learner behaviors (e.g., in MOOC platforms), thereby enabling more representative archetypes (Kizilcec et al., 2013; Breslow et al., 2013). Concurrently, HCI research introduced scalable techniques for persona generation by leveraging telemetry and click-stream data to establish archetypal learner segments (Zhang et al., 2016; Jung et al., 2018; Jansen et al., 2022; Salminen et al., 2022; Nielsen, 2019). However, despite methodological advances, traditional student persona generation frameworks remain constrained by static representations, reliance on manual curation, and limited capacity for large-scale, behaviorally consistent synthetic learner generation.

2.2 Student Persona Generation via LLMs

Initial LLM-based persona generation typically uses direct role prompting, e.g., instructing the model to `act as a student` or "write a learner profile" (Li et al., 2016; Zhang et al., 2018). More broadly, prompt design and prompt-based adaptation have been shown to substantially affect zero-shot behavior, controllability, and task transfer across NLP settings (Lu et al., 2023b; Liu et al., 2023; Lu et al., 2023a). Although highly scalable, this method frequently suffers from self-contradiction in one-shot long-context generation. (Li et al., 2024; Mündler et al., 2024). Subsequent work introduced schema-based controls (e.g., contradiction-aware rewriting, trait-consistency filters) (Song et al., 2020; Kim et al., 2020), and retrieval-augmented generation (RAG) to anchor persona attributes in external evidence (Lewis et al., 2020). Memory-augmented agent architectures preserve persona states across interactions for longitudinal coherence (Park et al., 2023). Reason-and-act pipelines coordinate writing, planning and tool use for multi-step synthesis (Yao et al., 2023). Controlled decoding methods (e.g., PPLM, GeDi) impose attribute filters for safer, more population-constrained outputs (Dathathri et al., 2020; Krause et al., 2021). Yet, two key gaps remain: lack of cohort-level distributional control and absence of education-theory-anchored validation, which hinder large-scale, theory-aligned and reliable student persona generation.

2.3 Student Persona Datasets

High-fidelity student simulation hinges on high-quality student persona data, yet such resources remain scarce (Wang et al., 2024). MathDial addresses this gap by pairing real teachers with LLMs acting as students, collecting tutoring dialogues to simulate learner misconceptions and scaffolding behavior (Macina et al., 2023), while Book2Dial generates cross-subject educational conversations from textbook materials to alleviate data scarcity (Wang et al., 2024). Other datasets approximate learner personas through dual-role simulation (Stasaski et al., 2020), classroom transcription (Suresh et al., 2022; Demszky and Hill, 2023), or real-world teacher–student chat logs (Caines et al., 2020). While large-scale educational logs are abundant, such as the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017), ASSISTments' online homework logs (Selent et al., 2016), the PSLC DataShop from learning sciences (Stamper et al., 2010), and EdNet's self-directed learning traces (Choi et al., 2020), they typically lack explicit schemas for learner personas. Student modeling efforts remain structurally rootless and difficult to validate in the absence of declarative constructs (Farooq et al., 2025; Tseng et al., 2024).

3 HACHIMI Framework

3.1 Problem Formulation

While prior work often treats persona generation as a byproduct of prompt engineering or dataset annotation, we explicitly define the task as Theory-Aligned and Distribution-Controllable Student Persona Generation (TAD-PG). This task requires producing a set of personas that are: (i) aligned with high-level educational objectives, (ii) internally coherent in discourse and traits, and (iii) matched to predefined population-level distributions.

3.2 Theory-Anchored Persona Schema

In designing student personas, we adopt *high-level education* as the guiding orientation: LLM-based educational models should go beyond answer accuracy and knowledge mastery to support value formation, personalized support, creativity development, and mental well-being (Kasneci et al., 2023; Durlak et al., 2011). These dimensions are treated as core educational outcomes in both Chinese curriculum standards under the *core-competencies–holistic development* paradigm (Lin, 2017) and the OECD vision of learner agency, creativity, and well-being (OECD, 2019).

To operationalize this vision at the student-persona level, we propose a theory-anchored persona schema that decomposes each persona into five complementary components based on the OECD Learning Compass (OECD, 2019): (1) **demographic & developmental status**, covering grade level and major moral-development phases (Kohlberg and Hersh, 1977); (2) **academic profile**, summarizing strong/weak subject clusters and an achievement tier; (3) **personality & value orientation**, aligned with value-education and character-formation constructs (Lapsley and Narvaez, 2006; Berkowitz and Bier, 2005); (4) **social relations & creativity**, capturing interaction patterns and creative problem-solving characteristics (Rubin et al., 2006; Runco, 2004; Beghetto and Kaufman, 2014); and (5) **mental health & well-being**, reflecting emotional functioning and support systems (Weare and Nind, 2011; Durlak et al., 2011). See Appendix B for design details. Together, these five components provide observable carriers that *instantiate* the four high-level educational capacities on the student side, spanning both relatively stable traits and interaction-driven states.

Within the TAD-PG task, the schema functions both as a structured constraint interface for theory-aligned, quota-controlled generation and as the state space for student agents, so that changes in values, personalization, creativity, and mental health can be read out as quantitative signals of high-level educational impact.

3.3 Multi-Agent Generation Architecture

We cast student persona generation as a *constrained optimization task* that reconciles LLM flexibility with the requirements of educational theory. The HACHIMI architecture employs a collaborative society of agents governed by three mechanisms, ensuring educational validity and structural consistency. Figure 1 gives an end-to-end overview; here we describe the three mechanisms, with stage-wise implementation details in Appendix C. We also provide the canonical prompt templates for all generation agents, together with the shared preamble, validator prompts, and responsible-field mapping, in Appendix D.

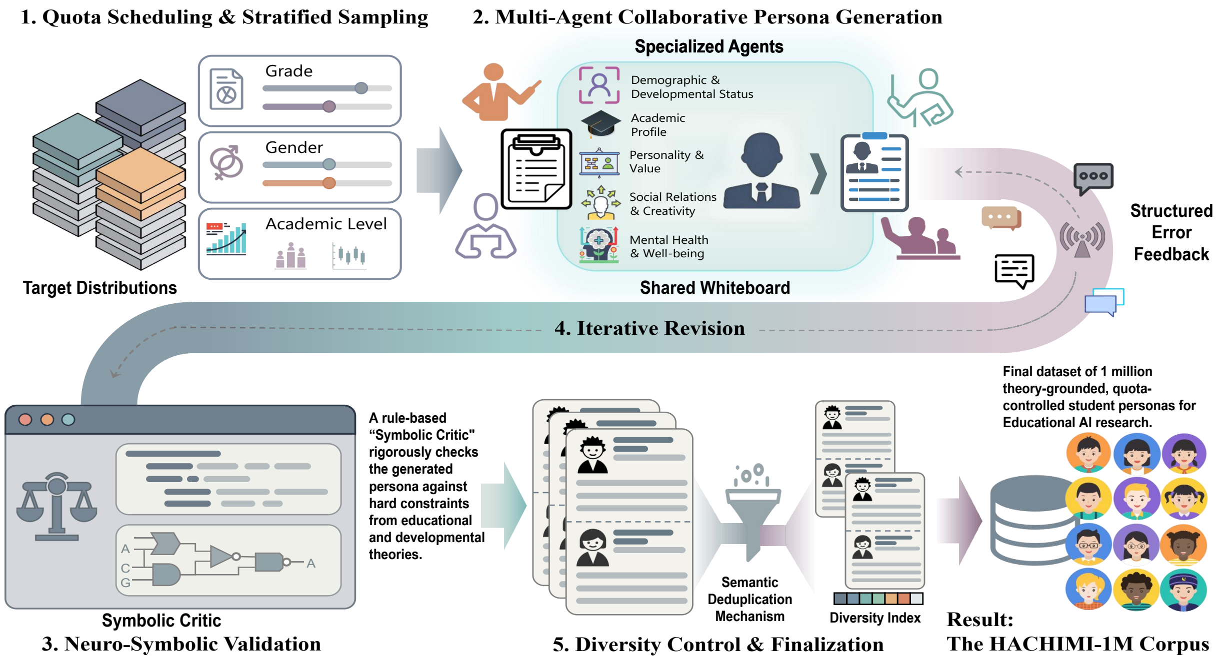

Figure 1: **HACHIMI pipeline overview.** From target distributions (grade/gender/academic level), steps (1)–(5) produce the HACHIMI-1M corpus.

Mechanism I: Modular Generation via Shared Whiteboard. Generating a *holistic student* in a single pass often leads to *intra-profile inconsistency* in long contexts, where different parts of the same persona may become self-contradictory or semantically misaligned (Li et al., 2024; Mündler et al., 2024). We therefore factorize personas into the five components in § 3.2. To avoid fragmentation across independently generated components, we anchor all agents to a **Shared Whiteboard** (Park et al., 2023), a dynamic context that lets agents condition sequentially on peers' intermediate states.

Mechanism II: Neuro-Symbolic Constraint Satisfaction. To mitigate stochastic LLM hallucinations, we adopt a **Propose-Validate-Revise** workflow inspired by iterative self-correction (Madaan et al., 2023). Neural agents generate *creative narratives*, while a rule-based Validator acts as a **Symbolic Critic** that rigorously checks drafts against educational axioms. Unlike simple heuristic filtering, we formalize theoretical alignment as satisfaction of logical predicates: demographic attributes are mapped to developmental stages (Piagetian (Piaget, 1964); Eriksonian (Erikson, 1963)) via *deterministic topology*, and trait descriptions are validated with set-theoretic logic. Upon violations, the system returns structured error signals to the relevant agents and iterates until all constraints are satisfied. We implement this critic as a rule-based Validator that checks the generated profile against an executable constraint set. The full executable rule set (R1–R15) is reported in Appendix E.

Mechanism III: Stratified Sampling and Diversity Control. To mitigate *mode collapse* (i.e., models converging to generic, averaged personas), we replace random sampling with **Stratified Sampling**. We enforce a uniform distribution over academic proficiency strata, ensuring balanced representation across student groups and serving as a **Conditional Variable** that propagates influence to downstream attributes (e.g., self-efficacy). To further maximize diversity, we apply **Semantic Deduplication** via Locality-Sensitive Hashing (LSH) (Charikar, 2002), mapping narratives into hash space and strictly removing semantically redundant duplicates to preserve a heterogeneous student population.

3.4 The HACHIMI-1M Corpus

As the primary artifact of our framework, we present the **HACHIMI-1M Corpus**, comprising **1,000,000** synthetic student personas spanning Grades 1 through 12. To the best of our knowledge, this represents the largest publicly available dataset of student profiles explicitly anchored in educational theory.

We include a persona example in Appendix G. All personas are generated with Qwen2.5-72B, and HACHIMI-1M can be produced at scale with ∼*3,200 H100 GPU-hour* compute (Appendix F).

Hybrid Semi-Structured Representation. Unlike traditional datasets comprised of interaction logs or purely unstructured text, HACHIMI-1M employs a **hybrid semi-structured format**. Each entry integrates two distinct data types:

1. **Categorical Labels:** Deterministic attributes for Piagetian and Eriksonian developmental stages and academic tiers. These serve as structured metadata for filtering and retrieval.

2. **Construct-driven Narratives:** Natural language descriptions for complex attributes drawn primarily from the *personality & value orientation*, *social relations & creativity*, and *mental health & well-being* components of our schema. These narratives are conditioned on specific psychological constructs, ensuring interpretability and theoretical consistency.

Distribution via Stratified Sampling. To ensure robust coverage across the student ability spectrum, we strictly employ a **Stratified Sampling** strategy. We enforce a uniform distribution (∼250,000 profiles per tier) across the four academic proficiency levels defined in §3.3. This approach effectively ensures that underrepresented groups, like struggling learners, are oversampled relative to their real-world frequency. This balanced distribution provides a standardized benchmark for evaluating the capabilities of educational AI systems. Offline descriptive summaries of the realized Grade × Academic Level and Grade × Gender distributions are reported in Appendix H.

4 Evaluation Methods

4.1 Evaluation Goals and Research Questions

Our evaluation aims to answer a central question: *to what extent do HACHIMI personas and persona-based student agents behave like real students at the group level while satisfying the TAD-PG task?*

We therefore focus on three complementary goals: (i) checking whether the generated persona collection satisfies the theory-anchored schema and quota targets, (ii) testing group-level consistency with the China Education Panel Survey (CEPS) (China Education Panel Survey (CEPS) Project, 2015), and (iii) investigating their transferability across broader geographical regions on the PISA 2022 dataset (OECD, 2023).

Based on these goals, we formulate the following research questions:

RQ1: Without relying on external data, does HACHIMI produce persona collections that respect the theory-anchored schema, match target quotas for grade, gender, academic achievement and psychological risk, and remain semantically diverse rather than collapsing into a few templates?

RQ2: When personas are instantiated as student agents in a CEPS Grade 8 shadow survey, do agent cohorts reproduce the item-wise relative differences across real-student cohorts?

RQ3: Do similar group-level consistency patterns hold on PISA 2022 constructs across world regions, suggesting that HACHIMI captures educational regularities rather than overfitting to CEPS?

4.2 Intrinsic Evaluation of TAD-PG Persona Collections

Before bringing in real student data, we first evaluate the HACHIMI persona collection itself with a lightweight offline evaluator. At the profile level, we distinguish *hard errors* (e.g., missing required fields, invalid fixed-label formats, or clear age–grade/developmental-stage violations) from *warnings* (e.g., partial dimension coverage, mild formatting problems, or softer consistency mismatches). We define **schema validity** as the fraction of personas that contain no hard errors, and we additionally report the warning rate. At the corpus level, we measure **quota satisfaction** by comparing the empirical distributions of grade, gender, academic level, and psychological risk against the scheduler targets; **diversity** by corpus-level lexical diversity indicators; and **redundancy** by approximate near-duplicate detection over long-text fields. All metric definitions, thresholds, and implementation details are reported in Appendix I.

4.3 Shared Protocol for External Group-level Evaluation

Both CEPS and PISA follow the same external evaluation logic: we first construct cohort-level reference patterns from real-student data, then instantiate HACHIMI personas as student agents, aggregate their shadow-survey responses by matched cohorts, and finally compare the two sides at the cohort level.

Persona-side instantiation and aggregation. On the HACHIMI side, personas are not grouped by subjective reading of free-form narratives. In-

Prompt Template for Shadow Survey

Role: You are a **student agent**. Answer strictly from the perspective of this student persona.
Persona: *$Persona*.
Questionnaire: *$Questionnaire*.
`## A list of CEPS/PISA-style items with IDs and discrete options.`
Instructions.
1. Stay fully in character and keep answers consistent with the persona.
2. For each item, choose exactly one option.
3. **Output a single JSON object** with integer codes.
Answer:

Figure 2: Immersive role-playing prompt template used for HACHIMI student agents when answering CEPS- and PISA-based shadow surveys.

stead, for each evaluation setting, we first filter the corpus to the required scope, then assign each eligible persona to the target cohort (s) using validated structured fields together with fixed parsing rules, and finally perform stratified quota sampling to obtain structurally matched agent populations. Each sampled persona is instantiated as a DeepSeek-V3.2-based student agent and prompted under the immersive role-playing scheme in Figure 2 to complete the corresponding shadow survey. The responses are mapped back to numeric codes and aggregated by cohort.

Shared consistency metrics. For any evaluated item or construct, the real-student side and the agent side each yield a cohort-level mean vector. We compare these vectors using Pearson correlation r to measure alignment in absolute group means (linear trend consistency), and Spearman correlation ρ to measure alignment in the relative ordering of cohorts (rank-order consistency). The two datasets differ in how cohorts and targets are defined, but they share this comparison protocol.

4.4 Group-level Consistency on CEPS (Chinese Grade 8)

We use the CEPS Grade 8 data as a nationally representative real-student reference and test whether HACHIMI agent cohorts reproduce the corresponding group-level regularities.

Real-student cohorts on CEPS. On the CEPS side, we focus on Grade 8 students and derive three stratification variables: academic achievement level (high / medium / low / poor), gender (M / F)[1], and psychological risk (high / low) from the CES-D depression scale in CEPS. Combining these three axes yields $4 \times 2 \times 2 = 16$ mutually exclusive cohorts. For each CEPS item selected for evaluation, we compute the mean response within each cohort, yielding a 16-dimensional vector of cohort means that serves as the real-world reference pattern for that item. Detailed preprocessing, variable construction, and cohort definitions are given in Appendix J.1.

Persona-side CEPS cohort assignment and CEPS-based shadow survey. On the persona side, we first retain only Grade 8 personas. Each retained persona is then deterministically assigned to one of the 16 CEPS-style cohorts by reading its validated `Academic Level` and `Gender` fields and by extracting the psychological-risk signal from the validated `Mental Health` field under fixed parsing rules that map the text to the same binary high/low risk split used for cohort matching. Only personas whose three stratification variables can be stably parsed are included in the CEPS analysis. We then perform stratified quota sampling and draw a fixed sample of 200 personas for each cohort. To obtain behaviour that is comparable to CEPS, we construct a shadow survey by selecting Grade 8 CEPS items that can be reasonably inferred from a textual persona. We retain attitudinal, perceptual, and self-reported behavioural items, and exclude purely factual items that cannot be inferred from the persona. The sampled personas are then instantiated and evaluated under the shared protocol in Section 4.3, producing agent-side cohort statistics in exactly the same format as the human data.

CEPS comparison target. Under this setup, each CEPS shadow-survey item yields one human-side and one agent-side 16-dimensional cohort-mean vector. We apply the shared consistency metrics above and report the resulting correlation distributions in § 5.2. Low-level implementation details remain in Appendix J.2.

4.5 Group-level Consistency on PISA 2022

We next investigate whether the CEPS-based consistency patterns transfer to an international assessment context on the PISA 2022 student dataset (OECD, 2023). The overall comparison protocol is the same as above, while the main dif-

[1]M/F follows the survey coding, used only for cohort matching; not intended to imply gender is binary.

ferences lie in the choice of constructs, regional grouping, and cohort definition.

Real-student constructs and cross-regional cohorts. On the PISA side, we select indices capturing students' motivation, affect, well-being, and creativity. Countries are grouped into five macro-regions (East Asia, Western Europe, Southern Europe, Latin America, Middle East), and within each region we form $4 \times 2 \times 2 = 16$ cohorts by crossing academic achievement quartiles, gender (M / F), and a binary psychological-risk indicator derived from the mental-health scale in PISA. For every region–construct pair, we compute unweighted cohort means, yielding a 16-dimensional cohort-mean vector as the real-student reference pattern. See Appendix K.1, K.2 for variable lists and thresholds.

PISA-based shadow survey. For each construct, we design a short PISA-based shadow survey by translating one or a few representative items into Chinese while preserving the original response scales. HACHIMI personas are then evaluated under the shared protocol in Section 4.3, with PISA-side grouping and matched cohort sampling, and responses are aggregated by region and cohort. See Appendix K.3 for item selection and coding.

PISA comparison target. For each region and construct, this procedure yields one human-side and one agent-side 16-dimensional cohort-mean vector. We apply Pearson and Spearman correlations and report the resulting cross-regional distributions in § 5.3; see Appendix K.3 for aggregation and implementation details.

5 Results

We report results from *inside-to-outside*. We first examine whether the generated persona collection itself satisfies the TAD-PG requirements (RQ1). We then move to external evaluation, first in a single national educational setting with CEPS (RQ2), and then in a cross-regional setting with PISA 2022 (RQ3). Finally, we compare HACHIMI against a protocol-matched one-shot baseline to isolate the gains brought by the generation framework itself.

5.1 Intrinsic Properties of TAD-PG Persona Collections

We begin with RQ1, which asks whether HACHIMI can generate persona collections that are structurally valid, quota-faithful, and semantically diverse before any comparison with real-student data. To answer this, we run the intrinsic evaluator (Appendix I) on a random, corpus-wide sub-corpus of $\sim$150,000 HACHIMI personas.

The results are uniformly strong. **Schema validity is near-perfect**: no sampled persona triggers any hard error, and only 0.06% receive soft warnings. At the corpus level, **quota targets are met almost exactly** (KL ≈ 0), indicating that the scheduler-controlled generation process remains tightly aligned with the intended marginal distributions. At the same time, **diversity remains high**: distinct-1/2 are approximately $0.40/0.83$, no SimHash near-duplicates are detected, and cross-component overlap remains low. In addition, *anchor alignment* between academic tier and the values / creativity / mental-health components is consistently strong, suggesting that the long-text descriptions track the intended academic-level conditioning in a stable and interpretable way.

Taken together, these findings provide a positive answer to **RQ1**: HACHIMI does not merely generate well-formed personas, but generates a corpus that simultaneously satisfies structural constraints, realizes the target quota design, and avoids semantic collapse. See Appendix L.1 for full results.

5.2 Group-level Consistency on CEPS

We next move to external evaluation in a single educational system and ask whether HACHIMI personas, once instantiated as student agents, can recover the group-level structure observed in real CEPS Grade 8 students. Following Section 4.4, we evaluate a set of CEPS targets that combine construct aggregates and focal items: six indices aggregated from related items (depression, parental strictness, teacher attention, misbehaviour, prosocial behaviour, and school bonding) and five stand-alone central items (aspiration, future confidence, parental expectations / pressure, and self-rated health). For each target, we correlate the 16-dimensional cohort-mean vectors between students and agents; see Appendix J.3 for definitions.

Figure 3 summarizes these correlations for CEPS constructs and focal items. The pattern is clear: **alignment is strongest for school-facing and academically grounded constructs**. Educational aspirations (`w2b18`) and parental achievement expectations (`w2a27`) are highly aligned (Spearman $\rho \geq 0.90$, Pearson $r \geq 0.86$), and aggregated teacher attention is similarly strong (Spearman $\rho \approx 0.90$, Pearson $r \approx 0.86$). Prosocial behaviour and future confidence reach moderate-to-high con-

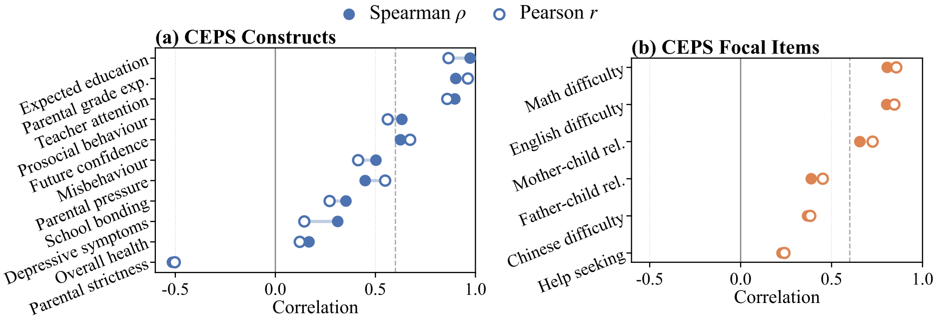

Figure 3: Pearson r and Spearman ρ between human and HACHIMI cohort means for each CEPS target.

sistency (Spearman $\rho \approx 0.63$), while misbehaviour frequency and parental-expectation pressure remain in the moderate range.

By contrast, constructs tied more closely to latent well-being and family dynamics are harder to recover from static personas. School bonding, depressive symptoms, self-rated health, and especially parental strictness show weak or even negative correlations, indicating that these dimensions are less directly inferable from the persona state used in the shadow-survey setting.

The same pattern appears at the item level. Perceived difficulty in mathematics (`w2b02`) and English (`w2b04`) is highly aligned (Spearman $\rho \approx 0.81/0.80$, Pearson $r \approx 0.86/0.85$), and mother–child relationship quality (`w2a23`) is also aligned (Spearman $\rho \approx 0.66$, Pearson $r \approx 0.73$). This suggests that the agents can recover cohort-level gradients in academic stress and some coarse-grained relational signals, even when alignment is not equally strong across all psychosocial constructs.

Overall, the CEPS results provide a positive answer to **RQ2**: when instantiated as student agents, HACHIMI personas recover a substantial portion of the group-level structure seen in real Grade 8 students. At the same time, they reveal a clear *fidelity gradient*: alignment is strongest for observable, school-facing constructs, and weaker for more latent well-being and family-related patterns.

5.3 Cross-regional Consistency on PISA 2022

We then ask whether the CEPS pattern generalizes beyond a single country and survey system. Using the protocol in Section 4.5, we evaluate human–agent consistency across five PISA 2022 regions.

Across regions, human–agent cohort correlations are decisively above chance for most constructs. Math-related and curiosity / growth constructs usually exceed 0.6 and often exceed 0.8: `MATHEFF` stays at $r > 0.95$ across all five regions,

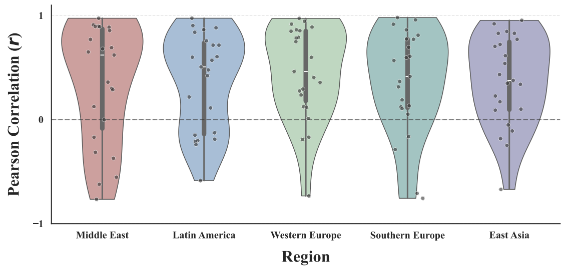

Figure 4: Distribution of Pearson correlations between human and agent group means on PISA 2022.

and `CURIOAGR` is similarly high ($r \gtrsim 0.85$). At a broad level, the correlations exhibit a tiered profile: very strong for math and curiosity-related constructs, moderate for classroom climate and belonging, and weak or negative for several well-being and workload indicators. This closely mirrors the CEPS pattern in § 5.2, suggesting that the same *fidelity gradient* persists across datasets rather than being specific to one survey source.

At the same time, PISA makes cross-regional structure more explicit. Math engagement and efficacy constructs show strong alignment in all regions (mostly $r > 0.7$), indicating that agents preserve the cohort ordering both *within* regions and *across* regions over the same 16 gender $\times$ academic-achievement $\times$ risk cohorts. Curiosity / growth constructs are aligned everywhere, although somewhat lower in East Asia and Southern Europe than in Latin America and the Middle East. By contrast, mental-health indices cluster around $r \approx 0$, workload and work–home balance variables are systematically negative, and some classroom-exposure constructs even flip sign across regions.

Figure 4 summarizes these regional distributions, while full per-construct results are reported in Appendix L.2. Overall, the PISA 2022 analysis supports **RQ3**: agent-based personas reproduce a stable cross-regional structure, with **strong alignment on math-related and curiosity constructs**, but

Table 1: Protocol-matched intrinsic comparison.

Metric	Baseline	HACHIMI	Δ
Hard error rate (↓)	12.03%	0.00%	-12.03
Warning rate (↓)	25.33%	0.82%	-24.51
Distinct-1 (↑)	0.2328	0.3285	$+0.0957$
Distinct-2 (↑)	0.4589	0.7893	$+0.3304$
Near-duplicate pairs (↓)	157	0	-157

much weaker alignment on latent well-being, workload, and some classroom-exposure variables.

5.4 Protocol-matched Baseline Comparison

Finally, to isolate the gains brought by the HACHIMI framework itself rather than by differences in evaluation protocol, we compare HACHIMI against a one-shot baseline under a matched protocol.

Compared with the one-shot baseline on matched 10,000 personas, HACHIMI **eliminates hard errors** and **improves diversity by a large margin**, as shown in Table 1. The baseline still exhibits substantial structural instability (12.03% hard errors; 25.33% warnings), whereas HACHIMI reduces these to 0.00% and 0.82%, respectively. At the same time, HACHIMI produces more diverse outputs (higher Distinct-1/2) and removes near-duplicate profiles entirely.

These gains also extend to **survey-grounded alignment on real-student data**. On CEPS, cohort-level correlations increase for teacher attention ($\Delta\rho = +0.132$, $\Delta r = +0.139$) and help-seeking ($\Delta\rho = +0.536$, $\Delta r = +0.641$). On PISA 2022, `MATHEASE` improves by $\Delta r = +0.27$–$+0.29$ (baseline $r = 0.45$–0.63), and `MATHEFF` improves by $\Delta r = +0.13$–$+0.14$ across regions. This indicates that HACHIMI's gains are not limited to better-formed personas in isolation, but translate into stronger external alignment under the same downstream evaluation pipeline.

Full baseline protocol and per-construct results are reported in Appendix M. For reproducibility, we report the experimental settings in Appendix A.

6 Discussion

This work shows that large-scale, theory-aligned and distribution-controllable student persona generation is feasible when treated as an explicit objective rather than a byproduct of prompting. Under the TAD-PG formulation, the HACHIMI multi-agent framework combines modular generation, neuro-symbolic validation and stratified diversity control to produce persona collections that satisfy schema and quota constraints while avoiding mode collapse. In this sense, HACHIMI-1M is **not just a dataset but an instantiated design space**: it demonstrates that educational theories, developmental taxonomies and quota schedulers can be integrated directly into the generation pipeline

Our empirical analyses reveal a structured pattern of group-level fidelity. On CEPS and PISA, persona-based agents most faithfully reproduce cohort differences in math-related efficacy, engagement, aspirations, teacher attention and curiosity/-growth constructs, while constructs tied to depressive symptoms, parental strictness, school bonding and workload balance are much harder to match. The cross-regional PISA results further show a tiered profile: robust alignment for math and curiosity, intermediate consistency for classroom climate and belonging, and weak or even negative correlations for well-being and workload indicators. This ***fidelity gradient*** suggests that *externally visible, school-facing aspects of student experience* are **more easily inferred** from static personas, whereas *latent mental health and family dynamics patterns* **remain underdetermined**.

These findings carry both opportunities and cautions for educational AI. HACHIMI-1M provides a standardized synthetic student population and a reusable testbed for cohort-matched benchmarking of educational LLMs and for social-science simulation, especially when real-student data are scarce or access-restricted; however, it should not be treated as a substitute for real-student evidence on subtle constructs such as well-being and family dynamics.

7 Conclusion

Theory-aligned, distribution-controllable student personas are hard to build with ad hoc *act-as-a-student* prompting. We presented HACHIMI, casting persona construction as TAD-PG with quota scheduling, multi-agent generation, neuro-symbolic validation, and diversity control. Across intrinsic tests and CEPS/PISA 2022, HACHIMI achieves near-perfect schema validity, accurate quotas, and stronger cohort-level alignment, strongest on school-facing math and curiosity/growth and weaker on latent well-being and family dynamics. Overall, HACHIMI positions student personas as structured, evaluable infrastructure for educational LLMs, enabling reliable agents and principled evaluation against large-scale survey data.

Limitations

While our findings are encouraging, several limitations warrant caution. First, our external validation relies on two large-scale survey datasets—CEPS Grade 8 and PISA 2022—which, although widely used, cover specific age ranges, curricula and socio-cultural contexts. HACHIMI personas and agents may behave differently when instantiated in other educational systems (e.g., vocational tracks, early childhood, or adult learners), and we do not claim cross-context generality beyond the settings examined in this work.

Second, our behavioural analysis focuses on group-level statistics over relatively short shadow surveys. We do not evaluate fine-grained, turn-by-turn classroom interactions, long-term learning trajectories, or micro-level causal effects of pedagogical interventions. A persona that reproduces cohort means on selected constructs may still exhibit unrealistic behaviour in extended dialogues, and our framework currently treats personas as static states rather than dynamically evolving learners. Third, we instantiate agents with a single family of LLM backbones and a specific prompting setup; different base models, decoding strategies, or role-playing prompts could lead to different degrees of alignment.

Finally, our approach remains constrained by the limitations of both LLMs and the underlying surveys. HACHIMI may inherit or amplify biases present in CEPS and PISA, and our theory-anchored schema necessarily simplifies complex constructs such as mental health, values, and family relations into a finite set of labels and narratives. We do not address fairness, privacy, or potential misuse in depth, and any real-world deployment of HACHIMI personas or agents should be accompanied by careful human oversight, ethical review, and, where possible, validation against up-to-date, context-specific student data.

Ethics Statement

This paper introduces HACHIMI, a framework for generating *synthetic* Grades 1–12 student personas and using persona-instantiated agents for cohort-level survey analyses. We emphasize that all personas in HACHIMI-1M are *fictional* and are not intended to represent, impersonate, or be linked to any real individual.

Use of real-student data. We use CEPS and PISA 2022 solely as *evaluation references* and conduct analyses at the *aggregated cohort level*. We follow the corresponding data-use terms and do not redistribute any restricted microdata. When constructing shadow surveys, we only report derived statistics and provide procedures such that replication relies on users' legitimate access to the original instruments and datasets.

Sensitive attributes and interpretation. Some labels (e.g., psychological-risk indicators derived from survey scales) are used only for *matched cohort analysis* and should not be interpreted as clinical diagnoses or used for individual-level screening. Similarly, M/F follows survey coding and is used only for cohort matching; it is not intended to imply that gender is binary. More broadly, the observed "fidelity gradient" indicates that certain latent or private constructs (e.g., well-being or family dynamics) are harder to infer reliably from static personas; outputs should therefore not be used for high-stakes decisions.

Bias, representativeness, and potential misuse. Synthetic personas may reflect biases in model priors and in the choice of schema, prompts, and constraints, and they may underrepresent minority or non-normative experiences. To mitigate these risks, we (i) make schema and control knobs explicit, (ii) report stratified evaluations across cohorts and regions, and (iii) recommend that future users validate conclusions against real-student evidence when feasible. We caution against using HACHIMI-1M for surveillance, automated profiling, or decision-making about real students; it is intended as a standardized testbed for benchmarking and simulation research.

Before release, we run automated screening to remove instances that contain obvious PII patterns (e.g., phone numbers, emails, URLs) and apply toxicity/offensiveness filters; flagged samples are discarded or regenerated.

Acknowledgments

This work is supported by the Shanghai Municipal Education Commission's Special Fund for Educational Large Models (93600-515100-25001).

References

Ronald A Beghetto and James C Kaufman. 2014. Classroom contexts for creativity. *High ability studies*, 25(1):53–69.

Marvin W Berkowitz and Melinda C Bier. 2005. What works in character education: A research-driven guide for educators.

Lori Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S. Stump, Andrew D. Ho, and Daniel T. Seaton. 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8:13–25.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher–student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer-Assisted Language Learning (NLP4CALL 2020)*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

John M. Carroll. 2000. Five reasons for scenario-based design. *Interacting with Computers*, 13(1):43–60.

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388.

China Education Panel Survey (CEPS) Project. 2015. China education panel survey (2014–2015 academic year) follow-up. Project ID: 61662993; Principal Investigator: Weidong Wang; Institution: Renmin University of China.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. *Lecture Notes in Computer Science*, pages 69–73.

Alan Cooper. 1999. The inmates are running the asylum. In Uwe Arend, Eckhard Eberleh, and Klaus Pitschke, editors, *Software-Ergonomie '99: Design von Informationswelten*, volume 53 of *Berichte des German Chapter of the ACM*, pages 17–17. Vieweg+Teubner Verlag, Wiesbaden.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR 2020)*.

Dorottya Demszky and Heather C. Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics.

Joseph A Durlak, Roger P Weissberg, Allison B Dymnicki, Rebecca D Taylor, and Kriston B Schellinger. 2011. The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, 82(1):405–432.

Erik H Erikson. 1963. *Childhood and society*, volume 445. Norton.

Ali Farooq, Amani Alabed, Pilira Stella Msefula, Reham Al Tamime, Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen. 2025. Representing groups of students as personas: A systematic review of persona creation, application, and trends in the educational domain. *Computers and Education Open*, 8:100242.

Bernard J. Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2022. *Data-Driven Personas*. Springer Nature, Cham.

Soon-gyo Jung, Joni Salminen, Haewoon Kwak, Jisun An, and Bernard J. Jansen. 2018. Automatic persona generation (APG): A rationale and demonstration. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*, pages 321–324, New York, NY, USA. Association for Computing Machinery.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*, pages 170–179, New York, NY, USA. Association for Computing Machinery.

Lawrence Kohlberg and Richard H Hersh. 1977. Moral development: A review of the theory. *Theory into practice*, 16(2):53–59.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952. Association for Computational Linguistics.

Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. *Scientific Data*, 4:170171.

Daniel K Lapsley and Darcia Narvaez. 2006. Character education. *Handbook of child psychology*, 4(1):696–749.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. Contradoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Chongde Lin. 2017. Research on core competencies of chinese students. *Psychological and Behavioral Research*, 15(2):145–154. (in Chinese).

Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023. Deeply coupled cross-modal prompt learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7957–7970, Toronto, Canada. Association for Computational Linguistics.

Jinghui Lu, Rui Zhao, Brian Mac Namee, and Fei Tan. 2023a. PUnifiedNER: A prompting-based unified NER system for diverse datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13327–13335.

Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. 2023b. What makes pre-trained language models better zero-shot learners? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2303, Toronto, Canada. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview.

Lene Nielsen. 2019. *Personas – User Focused Design*, 2 edition. Springer, London.

OECD. 2019. Oecd learning compass 2030: A series of concept notes. `https://www.oecd.org/education/2030-project/`. OECD Future of Education and Skills 2030.

OECD. 2023. Pisa 2022 database.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Jean Piaget. 1964. Cognitive development in children: Piaget. *Journal of research in science teaching*, 2(3):176–186.

John Pruitt and Jonathan Grudin. 2003. Personas: Practice and theory. In *Proceedings of the 2003 Conference on Designing for User Experiences (DUX '03)*, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Kenneth H Rubin, William Bukowski, and Jeffrey G Parker. 2006. Peer interactions, relationships, and groups. *Handbook of child psychology*, 3(5):619–700.

Mark A. Runco. 2004. Creativity. *Annual Review of Psychology*, 55:657–687.

Joni Salminen, Kathleen Wenyun Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2022. Use cases for design personas: A systematic review and new frontiers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, pages 1–21. Association for Computing Machinery.

Douglas Selent, Thanaporn Patikorn, and Neil T. Heffernan. 2016. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning at Scale*, pages 181–184, Edinburgh, UK. ACM.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online. Association for Computational Linguistics.

John Stamper, Kenneth R. Koedinger, Ryan S. J. d. Baker, Alida Skogsholm, Brett Leber, Jim Rankin, and Sandy Demi. 2010. Pslc datashop: A data analysis service for the learning science community. In *Intelligent Tutoring Systems*, volume 6095 of *Lecture Notes in Computer Science*, pages 455–455. Springer.

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. 2020. Cima: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631. Association for Computational Linguistics.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Book2dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731.

Katherine Weare and Melanie Nind. 2011. Mental health promotion and problem prevention in schools: what does the evidence say? *Health promotion international*, 26(suppl_1):i29–i69.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Shaoxiong Zhan, Yanlin Lai, Ziyu Lu, Dahua Lin, Ziqing Yang, and Fei Tan. 2026. MathSmith: Towards extremely hard mathematical reasoning by forging synthetic problems with a reinforced policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34602–34610.

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.

Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, pages 5350–5359, New York, NY, USA. Association for Computing Machinery.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2025. Simulating classroom education with llm-empowered agents. 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 10364–10379.

Yike Zhao, Simin Guo, Ziqing Yang, Shifan Han, Dahua Lin, and Fei Tan. 2025. More data or better data? a critical analysis of data selection and synthesis for mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 618–629, Suzhou (China). Association for Computational Linguistics.

A Implementation and Runtime Environment

For reproducibility, we summarise the main implementation and runtime settings below:

- **Programming language.** All data processing, persona scheduling, and agent-based evaluations (including CEPS and PISA shadow surveys) are implemented in Python 3.10.
- **Operating systems.** Experiments were conducted on standard workstation-class machines running recent versions of Linux and Windows.
- **Hardware.** All pipelines run on commodity CPUs without requiring specialised local GPUs; language model inference is performed via remote APIs rather than local deployment.
- **Python stack.** We rely on a standard scientific Python stack, including NumPy, pandas, SciPy, scikit-learn, and tqdm, plus common utilities for JSON/CSV I/O and logging.
- **LLM access.** Student agents are instantiated and queried through an OpenAI-compatible Python client connecting to the `deepseek-chat` model over HTTPS, with modest request-level concurrency to respect provider rate limits.

B High-level Education Capacities and Persona Schema

B.1 High-level Education Capacities: Behavioral Definitions

We operationalize "high-level education" along four capacities that guide both persona design and model evaluation:

Personalization. Through ongoing interaction rather than explicit labels alone, the model should implicitly construct and refine an internal student persona, dynamically diagnose the learner's knowledge state (including academic proficiency, value orientation, creativity and mental health), adapt to the student's cognitive style and pace, and provide targeted support that not only closes knowledge gaps but also promotes positive development in values, creativity and psychological well-being.

Value formation. When facing moral dilemmas, harmful content or value-laden questions, the model is expected to provide heuristic guidance, positive clarification and norm-consistent redirection, while upholding safety boundaries and avoiding value distortion.

Creativity support. The model should elicit and scaffold divergent thinking, pose Socratic questions, and promote distant associations between concepts, thereby improving the student's ability to generate, analyze and refine creative solutions.

Mental health support. The model should empathetically recognize and respond to students' emotions, provide support grounded in positive psychology, and respect safety boundaries and crisis-prevention principles, without stepping into non-professional diagnosis.

These behavioral definitions provide the target behaviors for teacher models and determine which aspects of each persona must be observable and editable for the TAD-PG task.

B.2 Mapping Capacities to Persona Components

The four capacities are not represented by a single field, but are distributed across five complementary persona components used in the TAD-PG task:

- **Demographic & developmental status** and **academic profile** primarily support *personalization*, by encoding grade, developmental stage and subject-specific strengths and weaknesses that the model can use to tailor instruction.
- **Personality & value orientation** provide the main observable substrate for *value formation*, while also informing how personalized guidance and mental-health responses should be framed.

- **Social relations & creativity** directly expose the student's social context and creative problem-solving patterns, enabling the model to exercise *creativity support* and to reason about peer/teacher/-family dynamics.

- **Mental health & well-being** offers a coarse but structured description of the student's emotional functioning and support systems, forming the primary basis for *mental health support* and for tracking changes in well-being.

In practical downstream applications, these component-level descriptors can serve as observable state variables for tracking whether teacher models move students toward desirable educational directions. In this paper, however, we do not evaluate such state transitions; instead, we use these components as static persona states for agent instantiation and group-level consistency analysis.

B.3 Persona Components in HACHIMI

HACHIMI instantiates the theory-anchored schema through five concrete components. Each component corresponds to one agent in the multi-agent architecture and is implemented as a structured set of fields.

B.3.1 Demographic & Developmental Status

This component is produced by the SCHOLAR agent and contains:

- **Name**: a full name.

- **Age** an integer between 6 and 18, loosely consistent with grade level.

- **Gender**: categorical.

- **Grade**: from primary grade one to senior high three.

- **Developmental stages**: a structured object with three keys—Piaget cognitive stage, Erikson psychosocial stage, and Kohlberg moral development stage, aligned with classical developmental theories (Kohlberg and Hersh, 1977).

- **Agent identifier**: a phonetic identifier following a constrained schema (1–2 syllables for surname, 1–3 for given name; each syllable is lowercase `pinyin` plus tone number; surname and given name separated by an underscore), ensuring uniqueness and downstream usability.

This component anchors the persona in a plausible age–grade band and provides the developmental frame needed to interpret value, creativity and mental-health descriptions.

B.3.2 Academic Profile

The ACADEMIC agent generates the academic profile:

- **Strong subjects**: a non-empty set of subjects drawn from clusters (STEM, humanities/social sciences, arts/PE, languages/biology), influenced by the sampling constraint on preferred subject cluster.

- **Weak subjects**: a non-empty set disjoint from strong subjects.

- **Achievement level**: a four-level categorical scale with fixed textual anchors:
  1. "High: top 10% in school",
  2. "Medium: top 10%–30%",
  3. "Low: top 30%–50%",
  4. "Poor: bottom 50%".

Achievement levels are used as hard anchors in quota scheduling and as conditioning signals for other agents, thus linking macro-level distribution control with micro-level content.

B.3.3 Personality & Value Orientation

The VALUES agent is responsible for:

- **Personality**: a short narrative description of personality traits (e.g., introversion/extraversion, conscientiousness, emotional stability), written in natural language.
- **Values**: a single-paragraph description that must explicitly cover seven dimensions, each with an interpretable level word, inspired by contemporary value-education frameworks:
    1. moral cultivation,
    2. physical and mental health,
    3. rule-of-law awareness,
    4. social responsibility,
    5. political identity,
    6. cultural literacy,
    7. family orientation.

Heuristics in the prompts and post-hoc filters prevent uniformly optimistic profiles by requiring a minimum count of "medium / low" level indicators for personas anchored to lower achievement tiers, thus tying value descriptions to the overall distributional design.

B.3.4 Social Relations & Creativity

The SOCIAL-CREATIVE agent produces:

- **Social relations**: a single paragraph (approximately 160–260 Chinese characters) describing peer, teacher and family interactions in a background–event–impact structure.
- **Creativity**: a single paragraph that combines eight dimension-specific judgments with a short "radar-style" summary, following creativity and CPS research (Runco, 2004; Beghetto and Kaufman, 2014):
    1. fluency,
    2. originality,
    3. flexibility,
    4. feasibility,
    5. problem finding,
    6. problem analysis,
    7. solution generation,
    8. solution refinement.

Each dimension must be associated with a level word and a brief justification. Internal consistency constraints (e.g., low feasibility cannot co-occur with very high solution generation) are enforced by the validator and light filters. As with values, creativity levels adapt to the academic-level anchor to suppress overly optimistic profiles.

B.3.5 Mental Health & Well-being

The HEALTH agent is in charge of the mental-health component:

- **Mental health**: a single paragraph that integrates:
    1. an overall summary of psychological functioning;
    2. at least two salient personality or temperament features;
    3. coarse indicators of overall mental state and subjective well-being (e.g., "overall mental status" and "happiness index");

4. risk descriptions for depression and anxiety (non-diagnostic, using educational language);
5. background stressors and protective factors (e.g., family, peers, school);
6. current supports and coping strategies.

Prompts explicitly require non-diagnostic language and coherence with the value dimension "physical and mental health", while filters prevent unrealistic combinations (e.g., very low achievement anchors with uniformly low risk and very high happiness).

B.4 Sampling Constraints and Distribution Control

For each persona, `_sampling-constraint` field encodes:

- target grade,
- gender,
- preferred subject cluster,
- target achievement level.

These constraints are generated by the quota scheduler to match pre-specified cohort-level distributions and are passed to all agents through the shared whiteboard. Agents must respect these constraints when producing their fields, and the validator checks alignment. As a result, the theory-anchored persona schema functions both as a conceptual bridge to high-level education capacities and as a concrete interface for distribution-controllable persona generation in the TAD-PG task.

C Overview of the HACHIMI Pipeline

Where to find the end-to-end overview. Figure 1 (main text, Section 3.3) gives an end-to-end overview of the HACHIMI pipeline, from distribution-aware sampling to the final HACHIMI-1M corpus for Theory-Aligned and Distribution-Controllable Persona Generation (TAD-PG). This appendix complements the main figure with a procedural summary (Algorithm 1) and stage-wise implementation details.

The pipeline consists of five tightly coupled stages:

1. Quota scheduling & stratified sampling. Given target distributions over *grade*, *gender*, and *academic level*, the scheduler allocates explicit quotas for each stratum and draws stratified samples of abstract "slots". Each slot encodes the macro-level variables required by the TAD-PG task (e.g., Grade 8, female, low-achievement, high-risk) and serves as a conditioning anchor for all subsequent agents. This stage enforces population-level control and guarantees coverage of under-represented groups such as struggling learners.

2. Multi-agent cooperative persona generation. For each scheduled slot, a society of specialized agents jointly constructs a holistic student persona on a shared whiteboard. Different agents are responsible for the four major components of the persona schema: *academic profile*, *personality & values*, *social relations & creativity*, and *mental health & well-being*. The shared whiteboard exposes the partial state (e.g., subject strengths, achievement tier, family background) so that later agents can condition on earlier decisions, preventing cross-component contradictions and mitigating intra-profile inconsistency in long-form generation.

3. Neuro-symbolic validation. The draft persona is then passed to a rule-based *Symbolic Critic*, which implements the neuro-symbolic constraints defined in the main text. The critic checks hard constraints derived from educational psychology and developmental theories (e.g., consistency between age and developmental stage, coherence between academic tier and self-efficacy, admissible combinations of risk factors). Instead of treating these checks as soft preferences, the validator encodes them as logical predicates over categorical labels and textual anchors extracted from the narrative fields.

Algorithm 1 HACHIMI pipeline.

Require: Target distributions over (grade, gender, academic level); persona schema $\mathcal{S}$; hard constraints $\mathcal{C}$; revision budget R
Ensure: Validated persona corpus $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$
$\mathcal{Q} \leftarrow \textsc{QuotaSchedule}(\text{targets})$
for all slot $s \sim \textsc{StratifiedSample}(\mathcal{Q})$ **do**
  $W \leftarrow \textsc{InitWhiteboard}(s)$
  $p \leftarrow \textsc{MultiAgentGenerate}(W, \mathcal{S})$
  for $t = 1$ to R **do**
    $(ok, err) \leftarrow \textsc{SymbolicCritic}(p, \mathcal{C})$
    if ok **then**
      break
    end if
    $p \leftarrow \textsc{Revise}(p, err, W)$
  end for
  if ok **then**
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{p\}$
  end if
end for
$\mathcal{D} \leftarrow \textsc{DiversityControl}(\mathcal{D})$
return $\mathcal{D}$

4. Iterative revision with structured error feedback. Whenever a violation is detected, the Symbolic Critic emits structured error messages that point to the offending components and the violated rules. These feedback signals are fed back to the relevant generators via the shared whiteboard, prompting targeted revision rather than unconditional regeneration. The loop continues until all hard constraints are satisfied or a small revision budget is exhausted, yielding a persona that satisfies the TAD-PG schema while preserving as much narrative richness as possible.

5. Diversity control & finalization. In the final stage, the system applies semantic diversity control over the pool of validated personas. A semantic deduplication mechanism (e.g., SimHash-based or other locality-sensitive hashing) flags near-duplicate narratives within the same stratum, and redundant entries are pruned or rewritten. Diversity indices at both token and construct levels are monitored to avoid mode collapse towards a few generic templates. The remaining personas are serialized into the hybrid semi-structured format described in the main paper, and aggregated into the HACHIMI-1M corpus as a theory-grounded, quota-controlled resource for downstream educational AI research.

D Agent Prompt Templates

This appendix provides the complete prompt templates for all agents in the HACHIMI multi-agent system. Each prompt box preserves the effective instruction template used in the implementation, while omitting low-level runtime wrappers such as API calls and logging.

D.1 Universal Agent Preamble (Shared by All Content Agents)

Universal Agent Preamble

```
You are a "student profile" production member collaborating with other agents. We use a "public
    whiteboard" to share drafts and discussions.

Rules (must follow):
- All output must be valid JSON objects, and only contain keys you are responsible for.
- Do not quote template phrases; use natural English; avoid empty cliches; avoid contradicting
    whiteboard draft.
- If asked to revise, only change keys you're responsible for; leave nothing empty; ensure logical
     consistency with other fields.
- Names should be in English; numbers and percentiles use English context (e.g., "top 10%").
- Do not output any extra explanatory text. Only output JSON.
- If "_Sampling Constraint" exists in the whiteboard, strictly follow "Grade", "Gender", "Strong
    Subject Preference", "Target Academic Level" and other requirements; if conflicts occur,
    follow sampling constraints and maintain overall consistency.
- Important: Every student has both strengths and weaknesses. Prohibited from writing all
    dimensions as "very good" or "clear advantage". When target academic level is "Mid/Low/Poor",
     your responsible fields must explicitly write some "Mid/Average/Weaker/Relatively Low/Low"
    descriptions to make the profile resemble a real person.

[Universal Output Hard Constraints] (All Agents must follow):
- You can only output one JSON object, not an array, and cannot add any explanatory text outside
    JSON.
- Absolutely prohibited from using ```json, ```, or other Markdown code block wrappers around
    output.
- The outermost layer of JSON can only contain keys this Agent is responsible for, must not add "
    id", "Student Info", "profile", or any other keys.
- Must not nest another meaningless wrapper object (e.g., structure like {"Student Info": {...}}
    is prohibited).
```

D.2 Enrollment & Development Agent

Responsible Fields: Name, Age, Gender, Grade, Developmental Stage, Agent Name
Task Mode: propose or revise

Enrollment & Development Agent Prompt

```
[AGENT_PREAMBLE]
[Sampling Constraint hint if present: Grade, Gender, Target Academic Level]

You are responsible for: ["Name", "Age", "Gender", "Grade", "Developmental Stage", "Agent Name"]
Task mode: {propose|revise}
Diversity seed: {seed}

Generation and Constraints (required):
- Age 6-18; Age must be an Arabic numeral integer; Grade must match Age (allow +/-1 year for skip/
    retention but must be consistent with other sections);
- Developmental Stage object must contain three keys: Piaget Cognitive Development Stage, Erikson
    Psychosocial Development Stage, Kohlberg Moral Development Stage;
- Agent Name format (multi-syllable support): Surname 1-2 syllables, Given Name 1-3 syllables, but
     2-syllable surname cannot pair with 3-syllable given name; each syllable is "lowercase
    pinyin+tone number(1-5)"; underscore between surname and given name, e.g., li3_gan4,
    jiang3_jie4shi2, ou1yang2_chen2fei1, yi4_yang2qian1xi3.

[Output Format Hard Constraints]:
- You can only output one JSON object, and top-level can only contain the following 6 keys:
  1) "Name" (string, English name)
  2) "Age" (integer, e.g., 12)
  3) "Gender" (string, can only be "Male" or "Female")
  4) "Grade" (string, e.g., "Grade 6", "Grade 7", "Grade 10")
  5) "Developmental Stage" (object, containing three subkeys)
  6) "Agent Name" (string, conforming to given regex)
- "Developmental Stage" must be an object and can only contain these three subkeys:
  - "Piaget Cognitive Development Stage"
  - "Erikson Psychosocial Development Stage"
  - "Kohlberg Moral Development Stage"
- Absolutely prohibited from adding "id", "Student Info", or other keys; no additional wrapper
    layer.
- Do not use ‘‘‘json or ‘‘‘ to wrap output.

[Qualified Example (strictly mimic structure, only change content)]:
[$case of JSON structure with Name, Age, Gender, Grade, Developmental Stage object with Piaget/
    Erikson/Kohlberg, Agent Name]

Please follow the above JSON structure, directly output the current student’s JSON object, no
    explanatory text.
```

D.3 Academic Profile Agent

Responsible Fields: Strong Subjects, Weak Subjects, Academic Level

Task Mode: propose or revise

Academic Profile Agent Prompt

```
[AGENT_PREAMBLE]

You are responsible for: ["Strong Subjects", "Weak Subjects", "Academic Level"]
Task mode: {propose|revise}
Diversity seed: {seed}

Field meanings and content requirements (required):
- "Strong Subjects": Non-empty array, each element is a subject name (e.g., "Chinese", "
    Mathematics"); no duplicates within array.
- "Weak Subjects": Non-empty array, each element is a subject name; set disjoint with "Strong
    Subjects".
- "Academic Level": Must strictly equal one of the following four strings:
  1) "High: Top 10% school ranking"
  2) "Mid: Top 10%-30% school ranking"
  3) "Low: Top 30%-50% school ranking"
  4) "Poor: Bottom 50% school ranking"
- [Subject preference hint if present]
- [Target Academic Level constraint if present: "This sample's 'Academic Level' must strictly
    equal: {level}"]

[Output Format Hard Constraints]:
- You can only output one JSON object, and top-level can only contain the following 3 keys:
  1) "Strong Subjects"
  2) "Weak Subjects"
  3) "Academic Level"
- These 3 keys must all appear, cannot be missing, cannot add any other keys.
- Absolutely prohibited from using "Student Info", "id", or other extra wrapper objects.
- Do not use ```json or ``` to wrap output.

[Qualified Example]:
[$case of JSON with Strong Subjects array, Weak Subjects array, and Academic Level string]

Please follow the above JSON structure, output only the JSON object itself.
```

D.4 Personality & Values Agent

Responsible Fields: Personality, Values

Task Mode: propose or revise

Adaptive Constraints (triggered when Target Academic Level is Mid/Low/Poor):

Personality & Values Agent: Adaptive Constraints

```
- [7-Dimension Mandatory Distribution (strongly bound to system filters, must satisfy)]
  - You must write both strengths and weaknesses across seven dimensions, cannot all be '
    Relatively High/Very High'.
  - Only allowed to use these level words: High / Relatively High / Upper-Mid / Mid / Relatively
    Low / Low.
  - When target academic level is 'Mid': Among seven dimensions, at least 1 dimension described
    with 'Mid/Relatively Low/Low'.
  - When target is 'Low': Among seven dimensions, at least 2 dimensions with 'Mid/Relatively Low/
    Low'.
  - When target is 'Poor': Among seven dimensions, at least 3 dimensions with 'Mid/Relatively Low/
    Low'.
  - For each 'Relatively High/Very High/Clear Advantage' description, pair with at least one 'Mid/
    Relatively Low/Low/Needs Improvement' description.
  - For each dimension written as 'Mid/Relatively Low/Low', provide brief justification.
```

D.5 Personality & Values Agent

Responsible Fields: Personality, Values
Task Mode: propose or revise

Personality & Values Agent Prompt

```
[AGENT_PREAMBLE]

You are responsible for: ["Personality", "Values"]
Task mode: {propose|revise}
Diversity seed: {seed}

Output format requirements (required):
- "Personality": One or several sentences of natural language, summarizing core personality traits
     (e.g., introverted/extroverted, responsibility, openness), maintaining education-scenario
    friendliness.
- "Values": Single-paragraph continuous natural language, no paragraphs, no lists or numbering.
  - Need to explicitly cover seven dimensions: Moral Character, Physical-Mental Health, Rule of
    Law, Social Responsibility, Political Identity, Cultural Literacy, Family Values;
  - Each dimension should have a recognizable level word (e.g., High/Relatively High/Upper-Mid/Mid
    /Relatively Low/Low), with brief justification.

[Conditional Adaptive Constraints, appended only when Target Academic Level is Mid/Low/Poor]:
- [7-Dimension Mandatory Distribution (strongly bound to system filters, must satisfy)]
  - You must write both strengths and weaknesses across seven dimensions, cannot all be '
    Relatively High/Very High'.
  - Only allowed to use these level words: High / Relatively High / Upper-Mid / Mid / Relatively
    Low / Low.
  - When target academic level is 'Mid': Among seven dimensions, at least 1 dimension described
    with 'Mid/Relatively Low/Low'.
  - When target is 'Low': Among seven dimensions, at least 2 dimensions with 'Mid/Relatively Low/
    Low'.
  - When target is 'Poor': Among seven dimensions, at least 3 dimensions with 'Mid/Relatively Low/
    Low'.
  - For each 'Relatively High/Very High/Clear Advantage' description, pair with at least one 'Mid/
    Relatively Low/Low/Needs Improvement' description.
  - For each dimension written as 'Mid/Relatively Low/Low', provide brief justification.

[Output Format Hard Constraints]:
- You can only output one JSON object, and top-level can only contain the following 2 keys:
  1) "Personality"
  2) "Values"
- Not allowed to have "Student Info", "id", "Evaluation", or any other top-level keys.
- "Values" must be single-paragraph text, no blank lines, no list symbols (e.g., "-", "1.", etc.)
    or Markdown in the middle.
- Do not use ```json or ``` to wrap output.

[Qualified Example]:
[$case of JSON with Personality description and Values single-paragraph covering 7 dimensions with
     level words]

Please follow the above JSON structure, output only JSON object.
```

D.6 Social & Creativity Agent

Responsible Fields: Social Relationships, Creativity
Task Mode: propose or revise

Social & Creativity Agent Prompt

```
[AGENT_PREAMBLE]

You are responsible for: ["Social Relationships", "Creativity"]
Task mode: {propose|revise}
Diversity seed: {seed}

Field and format requirements (required):
- "Social Relationships": Single paragraph (approx. 160-260 chars), narrate in "Background -> Key
    Event -> Impact" order, no line breaks or lists.
- "Creativity": Single-paragraph natural language, must include:
  - Evaluation of eight dimensions one by one: Fluency, Novelty, Flexibility, Feasibility, Problem
     Discovery, Problem Analysis, Proposing Solutions, Improving Solutions, each dimension with
    clear level word (High/Relatively High/Upper-Mid/Mid/Relatively Low/Low) and brief
    justification;
  - Ending with an overall "Radar Summary" comprehensive summary.
- If "Feasibility" relatively low or low, then "Proposing Solutions" no higher than mid-level,
    avoid self-contradiction.

[Conditional Adaptive Constraints, appended only when Target Academic Level is Mid/Low/Poor]:
- [8-Dimension Level Mandatory Distribution (strongly bound to filters)]
  - You need to first mentally assign a level word to each of 8 dimensions, only choose from 'High
    /Relatively High/Upper-Mid/Mid/Relatively Low/Low', then weave these 8 dimensions into a
    paragraph description.
  - Prohibited from writing all 8 dimensions as 'High/Relatively High', must have clear ups and
    downs.
  - When target academic level is 'Mid': Among 8 dimensions, at least 2 dimensions with 'Mid/
    Relatively Low/Low'.
  - When target is 'Low': Among 8 dimensions, at least 3 dimensions with 'Mid/Relatively Low/Low'.
  - When target is 'Poor': Among 8 dimensions, at least 4 dimensions with 'Mid/Relatively Low/Low
    '.
  - If 'Feasibility' written as 'Relatively Low/Low', then 'Proposing Solutions' level not allowed
     above 'Mid'.
  - The ending 'Radar Summary' must clearly point out which dimensions are strengths, which are
    obvious shortcomings. Must contain 'radar' or 'summary' word.

[Output Format Hard Constraints]:
- You can only output one JSON object, and top-level can only contain the following 2 keys:
  1) "Social Relationships"
  2) "Creativity"
- Not allowed to have "Student Info", "id", "Description", or any other keys.
- Both "Social Relationships" and "Creativity" must be single-paragraph text, cannot contain list
    symbols, numbering, Markdown, etc.
- Do not use ```json or ``` to wrap output.

[Qualified Example]:
[$case of JSON with Social Relationships paragraph and Creativity paragraph covering 8 dimensions
    with Radar Summary]

Please follow the above JSON structure, output only JSON object.
```

D.7 Mental Health Agent

Responsible Fields: Mental Health
Task Mode: propose or revise

Mental Health Agent Prompt

```
[AGENT_PREAMBLE]

You are responsible for: ["Mental Health"]
Task mode: {propose|revise}
Diversity seed: {seed}

Field and format requirements (required):
- "Mental Health": Single-paragraph natural language, no paragraphs or lists.
  Suggest naturally interspersing in the following order within paragraph:
  1) Overview of overall mental state;
  2) At least two personality traits related to psychological adaptation;
  3) Give clear level or degree descriptions for: Overall Mental State, Happiness Index,
     Depression Risk, Anxiety Risk;
  4) If no clear mental illness, include "Insufficient information or no significant symptoms" non
     -diagnostic description; if risks or tendencies exist, use "May have... tendency", "Mild...
     experience", "Recommend further assessment";
  5) Brief background story (e.g., academic pressure, interpersonal conflicts, family events);
  6) Current support and coping methods (family, teachers, peers, school resources).

[Conditional Adaptive Constraints, appended only when Target Academic Level is Low/Poor]:
- [Psychological Index Distribution (strongly bound to filters)]
  - Please explicitly give levels or degrees for four items in the text: Overall Mental State,
     Happiness Index, Depression Risk, Anxiety Risk.
  - When target academic level is 'Low/Poor':
    * Among Overall Mental State and Happiness Index, at least 1 item uses 'Mid/Average/Relatively
      Low/Below Average' mid-to-low descriptions, cannot both write 'Relatively High/Very High'.
    * Depression Risk and Anxiety Risk cannot both be written as 'Low/Very Low/Almost No Risk', at
      least 1 needs to show 'Mild/Some Possibility/Needs Attention' description.
  - Meanwhile maintain non-diagnostic tone for education scenarios: avoid directly using clinical
     diagnostic terms like 'Severe Depression/Bipolar/Need Hospitalization', use 'May have...
     tendency', 'Periodic Emotional Lows', 'Recommend Further Assessment' to describe.
  - When describing background stories and support systems, make readers feel: although some
     stress or distress exists currently, it can gradually improve through family, teachers,
     school resource cooperation.

[Output Format Hard Constraints]:
- You can only output one JSON object, and top-level can only contain 1 key:
  1) "Mental Health"
- Not allowed to have "Student Info", "id", "Evaluation", or other keys.
- "Mental Health" must be single-paragraph text, cannot contain blank lines, list symbols,
     Markdown code blocks.
- Do not use ```json or ``` to wrap output.

[Qualified Example]:
[$case of JSON with Mental Health single-paragraph containing overview, traits, 4 metrics,
     conditions, background, and support]

Please follow the above JSON structure, output only JSON object.
```

D.8 Validator Agent

Purpose: Comprehensive validation using all R1-R15 rules for fine-grained consistency checking
Input: Current whiteboard state (draft profile)

Validator Prompt

```
You are a 'Validator' agent. Please strictly review and provide structured revision tasks.
[AGENT_PREAMBLE]
You only output JSON, keys are issues and final_ready. Do not output extra text.

Reference Rules (required):
[Full R1-R15 rule block appended at runtime]

Output: issues: [{code, desc, owner, fields, hint}], final_ready: bool
```

Note. In the implementation, the placeholder "[Full R1–R15 rule block appended at runtime]" is replaced by the complete executable rule set reported in Table 3 in Appendix E.
Output Schema:

Validator Output Schema

```
{
  "issues": [
    {
      "code": "F1|F2|F3|F4",
      "desc": "Description of the issue",
      "owner": "Enrollment & Development|Academic Profile|Personality & Values|Social & Creativity
     |Mental Health",
      "fields": ["field_name"],
      "hint": "Suggestion for fix"
    }
  ],
  "final_ready": true|false
}
```

D.9 Field-to-Agent Responsibility Mapping

Table 2: Field-to-agent responsibility mapping in the HACHIMI multi-agent system.

Field	Responsible Agent
id	System
Name	Enrollment & Development
Age	Enrollment & Development
Gender	Enrollment & Development
Grade	Enrollment & Development
Developmental Stage	Enrollment & Development
Agent Name	Enrollment & Development
Strong Subjects	Academic Profile
Weak Subjects	Academic Profile
Academic Level	Academic Profile
Personality	Personality & Values
Values	Personality & Values
Social Relationships	Social & Creativity
Creativity	Social & Creativity
Mental Health	Mental Health

Table 3: Executable validator rules used in HACHIMI generation.

ID	Rule
R1	Age–grade norm: Grade 1–12 should approximately align with ages 6–18, allowing at most ± 1 year deviation.
R2	Developmental-stage plausibility: Piaget, Erikson, and Kohlberg stage labels must remain broadly compatible with the student's age band.
R3	`Strong Subjects` and `Weak Subjects` must both be non-empty and mutually disjoint.
R4	The eight creativity dimensions must show variation rather than identical levels; if feasibility is relatively low/low, proposing solutions cannot exceed a mid-level rating.
R5	If the values field presents stable positive physical/mental health, the mental-health field must not describe severe functional impairment or heavy clinical pathology.
R6	Agent-name regex: `^(?:[a-z]+[1-5]){1,2}_(?:[a-z]+[1-5]){1,3}$`.
R7	All required keys must be present and non-empty: `id`, `Name`, `Age`, `Strong Subjects`, `Weak Subjects`, `Grade`, `Personality`, `Social Relationships`, `Academic Level`, `Gender`, `Developmental Stage`, `Agent Name`, `Values`, `Creativity`, and `Mental Health`.
R8	The values paragraph must cover seven dimensions (Moral Character, Physical-Mental Health, Rule of Law, Social Responsibility, Political Identity, Cultural Literacy, Family Values), each with a locatable level expression.
R9	The creativity paragraph must include an overview, all eight dimensions (each with a level and brief rationale), and a radar-style summary.
R10	The mental-health paragraph must include: overview; at least two personality/adaptation traits; overall mental state; happiness index; depression/anxiety risk; a non-diagnostic risk/tendency statement; background context; and support/coping.
R11	Cross-field consistency: values, social, academic, and mental-health descriptions must be mutually supportive and not obviously contradictory.
R12	Non-diagnostic language: avoid heavy clinical wording (e.g., severe depression, bipolar disorder, medication, hospitalization); allow mild/tendency/situational/manageable/recommend consultation wording.
R13	Readability and anti-template constraint: reject mechanical repetition, laundry-list style text, or cases with missing level indicators/dimensions.
R14	`Values`, `Creativity`, and `Mental Health` must each be a single continuous paragraph, without lists, bullets, numbering, or multi-paragraph breaks.
R15	`Academic Level` must be one of four fixed labels only; otherwise the academic-profile owner must rewrite it using the strict four-choice format.

E Executable Validator Rules

We implement a two-stage validator. A fast validator performs low-cost structural screening, while a deep validator applies the full executable rule set (R1–R15) below during generation-time validation and revision, as shown in Table 3.

F Generation Efficiency and Key Speed-Related Settings

All personas are generated with Qwen2.5-72B(Yang et al., 2024; Team, 2024). We report compute in wall-clock time and GPU-hours.

Hardware and throughput. We run the end-to-end pipeline on 4×8 NVIDIA H100 GPUs (32 H100s total). The average throughput is ~10,000 personas/hour, and generating 1,000,000 personas takes ~100 hours, corresponding to ~3,200 GPU-hour in total.

Settings that affect runtime. We cap the Propose–Validate–Revise loop to at most 3 revision rounds. For semantic deduplication, we use SimHash with a near-duplicate threshold of Hamming distance ≤ 3; candidates within this threshold are discarded and resampled.

G A Persona Sample (Original in Chinese; Author-translated to English)

Note on language. The original persona record is written in Chinese for our internal generation and inspection workflow. For readership, we provide an translation into English below, in Table 4. All personas in our dataset are *synthetic* and use pseudonymous identifiers.

Table 4: One example persona (English translation from the original Chinese record).

Field	Content
Name (pseudonym)	Shihan Wang
Agent ID	`wang2_shi1han2`
Age / Gender	13 / Female
Grade	Grade 7 (Junior Year 1)
Academic standing	Low; bottom 50% in school ranking
Strengths	Art
Weaknesses	Mathematics; English; Physics
Developmental stage (theory-anchored)	**Piaget:** Formal operational stage; **Erikson:** Identity vs. role confusion; **Kohlberg:** Conventional level
Personality	Shihan is relatively introverted. She enjoys sharing her art creations with familiar classmates and expresses emotions in a delicate, nuanced way, but she tends to be reserved in group conversations. She can be patient when working on tasks; however, when facing academic challenges she easily falls into self-doubt. Her initiative is moderate and she has not yet formed stable self-discipline habits.
Values / civic literacy (inferred)	Her moral conduct is moderate: she respects classmates and occasionally helps close friends. Her physical and mental well-being is at a mid level, with occasional mood fluctuations triggered by academic stress or disappointing test results. Her rule-of-law awareness is relatively low: she has a shallow understanding of school rules and her sense of punctuality is not strong. Her sense of social responsibility is relatively low: she participates less in clubs and collective activities and focuses more on personal interests. Her political identity/engagement is relatively low: she rarely follows current affairs beyond the classroom. Her cultural literacy is moderate: she shows interest in traditional Chinese painting in art class but has limited exposure to other art forms. Her family orientation is high: when encountering learning difficulties, she mainly relies on her parents; family support plays an important role in her emotional life. (These judgments are inferred from her art-class performance, academic feedback, social participation, classroom behavior, and club participation.)
Social relationships	In Grade 7, Shihan maintains generally peaceful relationships with classmates. She often discusses her artwork with a few like-minded girls during art class or breaks, but appears timid with unfamiliar peers. After receiving recognition in a class art competition for her distinctive style, she became gradually more willing to join group work, though she still lacks a broad social circle. Compared with collective activities, she prefers working independently. When under pressure or facing academic difficulties, she tends to confide in family members or a small number of close friends. In events such as class dinners or sports days, she often joins the atmosphere when encouraged by nearby classmates, but her willingness to initiate conversations remains limited. Overall, her social support is moderate: relationships are stable but not wide-reaching.
Creativity profile	**Fluency:** medium; she can express personal themes relatively smoothly in drawing. **Novelty:** high; she often incorporates distinctive colors and topics. **Flexibility:** medium; she can try different art styles but rarely makes bold shifts in thinking. **Feasibility:** low; many creative ideas are difficult to implement into complete works. **Problem finding:** low outside art; she seldom proactively identifies problems in non-art subjects. **Problem analysis:** medium; she can analyze art-related bottlenecks but not deeply. **Solution proposing:** low; she lacks systematic plans when turning ideas into concrete artworks. **Solution improvement:** low; she shows limited initiative and methods to refine existing plans. Overall, she shows strong innovative awareness in art but weaker execution and sustained refinement; creativity is salient in one area but not well-balanced.
Mental health (non-clinical, descriptive)	Shihan is generally sensitive and tends to self-monitor, with noticeable emotional fluctuations under academic pressure and during group activities. She is quiet and prefers using art to express inner feelings, but she appears less confident in social settings and academic tasks. Overall, her well-being is slightly below medium; she often feels anxious about learning difficulties or poor performance, especially during exams and subject challenges. There is no evidence of severe mental disorder; she may occasionally show brief low mood and anxiety tendencies, both at a mid risk level, and the information is insufficient to diagnose any mental illness. Background: she has practiced art since primary school; her parents strongly support her artistic interests but sometimes place pressure on grades, making her particularly sensitive to criticism or failure. She mainly copes with stress through family/close-friend support and occasionally through drawing for self-regulation. Suggested support: strengthen positive family and teacher–student communication, encourage confidence building in a safe environment, gradually increase participation in group activities, help her recognize and express negative emotions, and cultivate self-acceptance.
Sampling constraints (for quota control)	Grade: Grade 7; Gender: Female; Preferred strength domain: art/music/sports; Target academic level: low (bottom 50%).

H Offline Distribution Summaries of HACHIMI-1M

As a supplement to the marginal quota diagnostics in Section 3.4, we report the offline count summaries of HACHIMI-1M by grade $\times$ academic level and grade $\times$ gender in Table 5.

Table 5: Offline summary of HACHIMI-1M by grade, academic level, and gender.

	Academic Level counts				Gender counts		
Grade	**High**	**Medium**	**Low**	**Poor**	**Male**	**Female**	**Total**
1	21,154	20,942	21,257	21,182	42,197	42,338	84,535
2	21,298	21,026	21,150	21,165	42,192	42,447	84,639
3	21,116	21,329	21,111	21,240	42,306	42,490	84,796
4	21,234	21,093	20,909	20,749	41,962	42,023	83,985
5	21,267	20,632	21,226	21,249	42,237	42,137	84,374
6	20,693	21,225	21,192	21,279	42,153	42,236	84,389
7	21,327	21,331	21,111	21,016	42,248	42,537	84,785
8	21,976	22,328	22,290	22,129	45,041	43,682	88,723
9	21,745	22,071	21,785	22,062	44,028	43,635	87,663
10	21,598	21,341	21,256	21,762	43,101	42,856	85,957
11	21,237	21,579	21,558	21,237	43,108	42,503	85,611
12	21,317	20,963	20,943	21,627	42,270	42,580	84,850

I Intrinsic Evaluation Details

I.1 Overview and Data Flow

We implement a standalone offline evaluator that operates on the merged persona corpus produced by HACHIMI (stored as a JSONL file where each line is a single persona). The evaluator does not modify or retrain any models; it only parses the generated personas and computes a set of post-hoc indicators. All reported intrinsic metrics in § 4.2 are derived from this evaluator.

I.2 Schema Validity and Theoretical Alignment

To assess whether the generated personas instantiate the TAD-PG task and theoretical constraints, we define a set of structural and theory-based checks:

- **Required field completeness.** We verify that all key fields are present and non-empty, covering identifiers, demographic fields (name, age, gender, grade), academic profile (achievement level, strong/weak subjects), developmental stages, agent identifier, and the three long-text components (values, creativity, mental health).

- **Academic profile well-formedness.** We enforce that the achievement level is one of four fixed textual anchors (high/ medium / low / poor; see § 3.4), and that strong and weak subject sets are non-empty lists with empty intersection.

- **Agent identifier format.** The "agent name" must follow a constrained `pinyin+tone` schema: surname and given name separated by an underscore, each part composed of one or more concatenated syllables of the form `[a-z]+[1-5]`, with the total syllable count corresponding to a plausible Chinese full name length.

- **Paragraph structure.** The three narrative components (values, creativity, mental health) must each be realized as a single, well-formed paragraph (no bullet lists or double line breaks). We also monitor paragraph character lengths and flag texts that are unusually short or long relative to predefined thresholds.

- **Value and creativity dimensions.** The values paragraph must explicitly mention the seven value dimensions (moral cultivation, physical and mental health, rule-of-law awareness, social responsibility, political identity, cultural literacy, family orientation) and include level expressions (e.g., high, medium, low). The creativity paragraph must cover eight creativity/CPS dimensions and contain a brief "radar-style" summary, with heuristics to flag logically inconsistent combinations (e.g., very low feasibility but unusually high solution-generation).

- **Mental health slots and non-diagnostic language.** The mental-health text is checked for the presence of key slots (overall mental status, happiness index, risk descriptions, stressors, supports, coping strategies) and for adherence to non-diagnostic educational language. We additionally flag combinations where mental-health descriptions are incompatible with the value dimension "physical and mental health" (e.g., "excellent health" paired with "severe depression").

- **Age–grade and developmental-stage consistency.** We apply coarse age–grade consistency checks using expected age ranges for each grade (primary, lower secondary, upper secondary). We also check whether Piagetian, Eriksonian and Kohlbergian stages are roughly compatible with the student's age band, and flag obvious violations (e.g., pre-operational stage assigned to a high-school student).

For each persona, the evaluator records violations under a concrete taxonomy of *errors* (hard schema violations) and *warnings* (soft theoretical mismatches):

- **Errors (hard schema violations).** A persona is marked as having an error if any of the following holds:
  - *Missing or empty required field:* any of the core slots is absent or empty (identifiers, age, gender, grade, academic level, strong/weak subject lists, developmental stages, agent identifier, or any of the three long-text components).
  - *Ill-formed academic profile:* the achievement level does not match one of the four allowed anchors (high / medium / low / poor), or the strong/weak subject lists are empty or have non-empty intersection.
  - *Malformed agent identifier:* the agent name violates the constrained `pinyin+tone` pattern (surname_givenname, with syllables of the form `[a-z]+[1-5]` and a plausible total syllable count).
  - *Invalid paragraph format:* any of the values, creativity, or mental-health components is not realized as a single paragraph (e.g., list-like formatting, multiple blank lines) and thus does not satisfy the schema.
  - *Missing theory-mandated dimensions:* the values text fails to cover all seven value dimensions with explicit level words, the creativity text omits one or more of the eight CPS dimensions, or the mental-health text lacks core slots such as overall status, risk/stressors, supports, or coping strategies.
  - *Age–grade / stage inconsistency:* the age–grade combination falls outside the allowed ranges for compulsory schooling, or the Piagetian / Eriksonian / Kohlbergian stages are incompatible with the age band in a way that cannot be justified as borderline (e.g., pre-operational stage assigned to an upper-secondary student).

- **Warnings (soft theoretical mismatches).** A persona is marked as having a warning if it passes the hard schema checks but triggers any of the following softer heuristics:
  - *Atypical paragraph length:* values, creativity, or mental-health paragraphs are unusually short or long relative to target ranges (flagging potential under-specification or verbosity).
  - *Partial dimension coverage:* some but not all value or creativity dimensions are mentioned, or some level words are vague or missing, while the overall structure remains acceptable.
  - *Creativity profile inconsistencies:* "radar-style" summaries contain mild contradictions (e.g., very low feasibility paired with extremely high overall CPS score) that are not severe enough to reject the persona.

- *Mental-health / value mismatches:* the mental-health text uses educationally undesirable diagnostic labels (e.g., clinical disorder names) or shows tension with the "physical and mental health" value dimension (e.g., "excellent health" co-occurring with strong negative symptom descriptions).
- *Borderline developmental mismatch:* the age band and developmental stages are slightly misaligned (e.g., a younger-than-typical student assigned to a later stage), but still within a borderline interpretable range.

At the corpus level, we report the proportion of personas with at least one error or warning, as well as summary statistics such as paragraph-length distributions. In addition, we construct simple text-derived "alignment scores" by mapping level words in the values, creativity, and mental-health texts to polarity weights and computing Pearson correlations between these scores and the four-level academic-achievement anchor; these correlations quantify how consistently textual descriptions track the intended academic tiers.

I.3 Distributional Alignment with Target Quotas

To evaluate the "distribution-controllable" aspect of TAD-PG, we compare the empirical distributions of key stratification variables against target quotas or generation plans:

- **Marginal distributions.** We compute empirical frequency distributions over grade, gender and academic-achievement level for the entire corpus and treat them as empirical distributions $P(\cdot)$. We compare these to pre-specified target distributions $Q(\cdot)$ (e.g., uniform over grades, balanced gender, uniform over the four academic tiers) using absolute deviation and Kullback–Leibler (KL) divergence $\mathrm{KL}(P\|Q)$ with simple Laplace smoothing.
- **Plan-based alignment (optional).** When a generation schedule is available (e.g., a JSON plan recording the intended grade, gender and academic level for each scheduled persona), we compute per-field *match rates* between generated and planned labels, as well as KL divergence between the empirical corpus distributions and the plan-induced distributions. This directly measures how tightly HACHIMI follows a concrete quota schedule when one is used.

These distributional diagnostics are summarized in the evaluation report as overall KL divergence values for each variable and, when applicable, as plan-alignment metrics. They provide a corpus-level check that HACHIMI does not drift away from the intended grade, gender and achievement landscape when scaling up to large persona populations.

I.4 Diversity and Redundancy Control

To ensure that the persona collection covers a wide variety of student types rather than collapsing into a few templates, we compute several diversity and redundancy indicators on concatenated persona texts:

- **Distinct-n.** We compute character-level Distinct-1 and Distinct-2 at the per-sample level (ratio of unique n-grams to all n-grams, averaged across personas) over a concatenation of key narrative fields. Higher Distinct-n values indicate richer lexical and phrasal variety.
- **SimHash near-duplicate detection.** For each persona, we compute a 64-bit SimHash over character trigrams of the concatenated long texts. We then approximate the distribution of pairwise Hamming distances via random sampling and record the number of persona pairs whose distance falls below a small threshold (i.e., near-duplicates). This provides both a global sense of redundancy and a concrete list of highly similar profile pairs.
- **Cross-component similarity.** We compute character-level Jaccard similarity between the values, creativity and mental-health paragraphs for each persona and flag cases where any pair exceeds a high threshold (e.g., 0.8). These "template-like" samples are counted as potential overuse of shared boilerplate across components.

The evaluator reports corpus-level Distinct-n values, summary statistics of SimHash Hamming distances, the count of near-duplicate pairs, and the number of template-like samples with excessive cross-component similarity. Collectively, these metrics characterize how well HACHIMI balances structural constraints with semantic variety.

I.5 Local Coherence and Contradiction Checks

Finally, we perform basic local-coherence checks to guard against salient contradictions within individual personas. These checks are primarily rule-based:

- **Hard contradictions.** We search for patterns where different fields convey mutually incompatible information, such as: extremely high academic achievement paired with long-term, severe perceived difficulty in the same subject; very high physical and mental health ratings co-occurring with explicit descriptions of extreme psychological distress; or age–grade combinations that are strongly implausible.

- **Slot-level consistency.** Within the mental-health component, we cross-check the overall mental status, happiness index and risk descriptions for depression/anxiety to ensure that severity terms (e.g., mild, moderate, severe) are not obviously inconsistent with the qualitative description of functioning and support.

- **Subject-mention coverage.** As a weak form of coherence between structured and narrative fields, we measure the proportion of personas for which strong/weak subjects explicitly appear in the long-text descriptions (values, social relations, creativity, mental health), indicating that narrative content meaningfully reflects the structured academic profile.

At the corpus level, we summarize the fraction of personas exhibiting any hard contradiction, the subject-mention coverage rates, and the distribution of errors and warnings per persona. We also identify the top-k most problematic personas (with the largest number of errors/warnings) to facilitate manual inspection. These diagnostics serve as a sanity check that TAD-PG constraints are not only satisfied at the field level, but also reflected in self-consistent narratives.

J Details of the CEPS-based Consistency Evaluation

J.1 Data Preprocessing and Cohort Construction

This section documents how we preprocess CEPS Grade 8 data and construct the stratified cohorts used as human baselines in § 4.4.

Raw data and basic filters. We use the 2014–2015 wave of the China Education Panel Survey (CEPS) student questionnaire and retain only Grade 8 students.[2] We drop records with missing values in any of the key fields needed for cohort construction (core subject scores, gender and depression scale items), resulting in N_{CEPS} valid Grade 8 cases.

Gender and gender imputation. For gender, we use the official CEPS gender variable and treat it as binary (M/F). A small fraction of cases lack this field; for these, we perform gender imputation based on mutually exclusive puberty indicators (e.g., age at first menstruation vs. age at first nocturnal emission). Concretely, we assign "F" when only the menarche item is non-missing and "M" when only the nocturnal-emission item is non-missing; records with ambiguous or missing puberty information are dropped.

[2]We rely on the official grade indicator provided in the CEPS documentation; implementation-wise this is a simple filter on the grade field.

Academic achievement level. We use the raw scores of Chinese, mathematics and English in CEPS Grade 8 (denoted s_{chn}, s_{mat} and s_{eng}). Each subject score is first standardised within the Grade 8 population,

$$z_{\text{chn}} = \frac{s_{\text{chn}} - \mu_{\text{chn}}}{\sigma_{\text{chn}}} \tag{1}$$

$$z_{\text{mat}} = \frac{s_{\text{mat}} - \mu_{\text{mat}}}{\sigma_{\text{mat}}} \tag{2}$$

$$z_{\text{eng}} = \frac{s_{\text{eng}} - \mu_{\text{eng}}}{\sigma_{\text{eng}}} \tag{3}$$

We then form a total achievement index

$$z_{\text{total}} = z_{\text{chn}} + z_{\text{mat}} + z_{\text{eng}}. \tag{4}$$

The distribution of z_{total} is sliced into four tiers by empirical percentiles:

- *High*: top 10% of z_{total};
- *Meduim*: 10–30%;
- *Low*: 30–50%;
- *Poor*: bottom 50%.

These four tiers are used as the academic-level labels in § 4.4.

Psychological risk from CES-D. To characterise psychological risk, we use the CES-D short depression scale embedded in CEPS. Let $d_1, \ldots, d_{10}$ denote the ten item responses (coded on the original CEPS Likert scale). We compute a simple sum score

$$\text{CESD} = \sum_{k=1}^{10} d_k. \tag{5}$$

We then define "high" vs. "low" psychological risk via the empirical 75th percentile of CESD within Grade 8:

- *High-risk*: CESD $\geq$ 75th percentile;
- *Low-risk*: CESD $<$ 75th percentile.

We do not apply any further weighting or clinical cutoffs, as our goal is to obtain a stable rank-based risk stratification rather than a diagnostic label.

Resulting stratified cohorts. Combining the four academic levels, two gender categories and two psychological risk levels yields $4 \times 2 \times 2 = 16$ mutually exclusive cohorts. Each Grade 8 student is assigned a unique triplet (academic level, gender, risk), and group-wise statistics (e.g., item means) are computed by taking unweighted averages within each cohort.

J.2 Consistency Metrics and Implementation Details

This section details how we construct the human and agent cohort-level statistics and compute the consistency metrics reported in § 4.4.

Item selection and coding. From the CEPS Grade 8 student questionnaire we select a subset of items that satisfy two criteria: (i) they reflect perceptions, attitudes or self-reported behaviours that can reasonably be inferred from a textual persona (e.g., perceived difficulty in mathematics, perceived parental expectations, class climate, prosocial or problem behaviours); and (ii) they have well-defined discrete response options with stable coding schemes. Purely factual items that cannot be inferred from the persona (such as height or weight) are excluded. We keep the original item IDs and numeric codes, and re-use them on the agent side so that human and agent responses are aligned at the code level.

Human cohort means. For each selected CEPS item j and each of the 16 cohorts $g \in \{1, \ldots, 16\}$, we compute the unweighted mean

$$\mu_{g,j}^{(H)} = \frac{1}{|S_{g,j}|} \sum_{i \in S_{g,j}} x_{i,j} \tag{6}$$

where $x_{i,j}$ is the numeric response of student i to item j, and $S_{g,j}$ is the set of students in cohort g with non-missing responses on j. These 16-dimensional vectors

$$\mu_j^{(H)} = \left(\mu_{1,j}^{(H)}, \ldots, \mu_{16,j}^{(H)}\right) \tag{7}$$

are treated as the human reference patterns for item j.

Agent cohort means. On the agent side, each sampled persona is instantiated as a student agent and prompted with the CEPS-based shadow survey (Figure 2). The LLM outputs are constrained to a JSON object whose keys are CEPS item IDs and whose values are the chosen option codes. After basic validation, we assign each agent to one of the 16 cohorts (using the academic level, gender and psychological risk labels embedded in the persona) and compute cohort means in exactly the same way as for humans:

$$\mu_{g,j}^{(A)} = \frac{1}{|T_{g,j}|} \sum_{a \in T_{g,j}} y_{a,j} \tag{8}$$

where $y_{a,j}$ is the numeric response of agent a to item j, and $T_{g,j}$ is the set of agents in cohort g with valid responses on j. This yields agent-side vectors

$$\mu_j^{(A)} = \left(\mu_{1,j}^{(A)}, \ldots, \mu_{16,j}^{(A)}\right) \tag{9}$$

for each item j.

Pearson and Spearman correlations. For each item j, we compute two scalar consistency metrics between humans and agents:

- **Pearson correlation** r_j between $\mu_j^{(H)}$ and $\mu_j^{(A)}$, capturing *linear trend consistency* in absolute cohort means;
- **Spearman rank correlation** ρ_j between the same vectors, capturing *rank-order consistency* in the relative ordering of cohorts.

Ties in cohort means are handled using average ranks in the Spearman computation. Items for which fewer than two cohorts have non-missing means on either side are excluded from correlation analysis.

Summary statistics and visualisation. In the main text we summarise the distribution of $\{r_j\}$ and $\{\rho_j\}$ across all evaluated items (e.g., mean, standard deviation and selected quantiles), and provide representative visualisations: scatter plots of human vs. agent cohort means for individual items, and grouped bar charts comparing the 16 cohort means for selected items that illustrate typical agreement or disagreement patterns. All statistics are computed in Python using standard scientific libraries (e.g., `pandas`, `numpy` and `scipy`).

J.3 CEPS Constructs and Aggregation Rules

For the CEPS-based consistency analysis in § 5.2, we select a focused subset of Grade 8 questionnaire items and group them into a small number of higher-level constructs. All raw items follow the original CEPS coding, and we use simple additive or averaging rules to obtain construct scores at the individual level before computing cohort means. For compactness, we denote the construct scores by $D(i)$, $S_{\text{parent}}(i)$, $T_{\text{att}}(i)$, $M(i)$, $P_{\text{prosocial}}(i)$ and $B_{\text{school}}(i)$ for the six constructs below.

Depressive symptoms. The construct `Construct_Depression` summarises psychological distress based on the ten C25 items (`w2c2501--10`), which ask how often students experience symptoms such as sadness, worry, sleep problems or loss of appetite. Each item is coded on a 1–5 frequency scale ("never" to "always"). We compute the construct score as the row-wise sum:

$$D(i) = \sum_{j=1}^{10} \texttt{w2c25}j(i), \tag{10}$$

with higher values indicating higher psychological risk.

Parental strictness. The construct `Construct_Parental_Strictness` captures how strictly parents regulate academic work and media use, using A20-1 (`w2a2001`, homework supervision) and A20-5 (`w2a2005`, internet supervision), both on a 1–3 scale. We define

$$S_{\text{parent}}(i) = \frac{\texttt{w2a2001}(i) + \texttt{w2a2005}(i)}{2}. \tag{11}$$

Teacher attention. The construct `Construct_Teacher_Attention_Avg` aggregates perceived teacher attention in mathematics, Chinese and English (B5-1 / `w2b0501`, B5-2 / `w2b0502`, B5-3 / `w2b0503`), each coded 1–4. We compute

$$T_{\text{att}}(i) = \frac{\texttt{w2b0501}(i) + \texttt{w2b0502}(i) + \texttt{w2b0503}(i)}{3}. \tag{12}$$

Misbehaviour. The construct `Construct_Misbehavior` captures externalising problem behaviours using D2-1 (`w2d0201`, swearing) and D2-3 (`w2d0203`, fighting), both coded 1–5. We define

$$M(i) = \frac{\texttt{w2d0201}(i) + \texttt{w2d0203}(i)}{2}, \tag{13}$$

with higher values indicating more frequent misbehaviour.

Prosocial behaviour. The construct `Construct_Prosocial` reflects prosocial tendencies using D1-1 (`w2d0101`, helping others) and D1-2 (`w2d0102`, following rules), again on a 1–5 scale:

$$P_{\text{prosocial}}(i) = \frac{\texttt{w2d0101}(i) + \texttt{w2d0102}(i)}{2}. \tag{14}$$

School bonding. The construct `Construct_School_Bonding` captures students' sense of connection to their class and school via B6-6 (`w2b0606`, class climate) and B6-7 (`w2b0607`, participation in activities), both coded 1–4. We compute

$$B_{\text{school}}(i) = \frac{\texttt{w2b0606}(i) + \texttt{w2b0607}(i)}{2}. \tag{15}$$

K Details of the PISA 2022-based Consistency Evaluation

K.1 PISA 2022 Variable Selection and Preprocessing

We base our PISA-side analysis on the public PISA 2022 student questionnaire data (OECD, 2023). To obtain construct scores that are comparable to our HACHIMI agents, we proceed in three steps: variable selection, basic filtering, and standardisation.

Construct selection. We focus on a set of psycho-social and learning-related OECD student-questionnaire indices (WLE) that reflect students' motivation, affect, well-being, and creativity. Concretely, we include:

- mathematics self-efficacy / efficacy indices (e.g., `MATHEFF`, `MATHEF21`);
- mathematics anxiety (e.g., `ANXMAT`);

- sense of belonging at school (e.g., BELONG);
- life satisfaction (e.g., LIFESAT);
- psychosomatic symptoms / psychological distress (e.g., PSYCHSYM);
- creative self-efficacy and creativity/openness to intellect (e.g., CREATEFF, CREATOP);
- social connections (ease of communication about worries and concerns) (e.g., SOCCON).

All selected variables are official PISA indices constructed by the OECD from multiple questionnaire items (reported as WLE scale scores). We use these released index scores directly (without re-deriving them from item-level responses) and only apply basic filtering and standardisation to align their scale and comparability with the agent-side construct scores.

For the purpose of defining achievement-based cohorts (Appendix K.2), we additionally use the full set of plausible values for mathematics, reading, science, and mathematics content/process subscales (e.g., PV1MATH--PV10MATH, PV1READ--PV10READ, etc.). We aggregate these into a composite achievement score as described below.

Sample filtering. We retain only students with valid values on: (i) the selected psycho-social constructs, (ii) the achievement plausible-value variables used to construct the composite score, and (iii) basic demographics (country, gender). Students with missing or invalid codes on any of these fields are dropped. No survey weights are applied, as our goal is to compare cohort-level patterns rather than to produce nationally representative estimates.

Normalisation. For descriptive reporting and cross-region comparability, we z-normalise each psycho-social construct separately within the pooled PISA sample:

$$z_{i,c} = \frac{x_{i,c} - \mu_c}{\sigma_c}, \tag{16}$$

where $x_{i,c}$ is the raw score of student i on construct c, and μ_c, σ_c are the mean and standard deviation of c across all retained students. For grouping and correlation analyses, we use these normalised scores but keep the original directionality (i.e., higher values always mean more of the underlying construct; reverse-coded scales are flipped where necessary).

K.2 Region and Cohort Construction

We next define the cross-regional cohorts used in § 4.5. The construction parallels the CEPS cohorts but operates on PISA countries and constructs.

Macro-region mapping. Each participating country or economy is mapped to one of a small number of macro-regions based on geographical and cultural proximity (e.g., East Asia, Western Europe, Southern Europe, Latin America, Middle East). Let $\mathcal{R}$ denote the set of regions, and let region$(i) \in \mathcal{R}$ be the region of student i. The exact country–region mapping table is provided in the project repository and omitted here for space.

Composite achievement quartiles. Let $\mathcal{V}_{\text{ach}}$ denote the set of plausible-value columns we use to characterise overall academic achievement (mathematics, reading, science, and mathematics content/process subscales; see Appendix K.1 for a description of the domains covered). For each student i, we first compute a composite achievement score by averaging all plausible values in this set:

$$\text{ACHV\_TOTAL}(i) = \frac{1}{|\mathcal{V}_{\text{ach}}|} \sum_{v \in \mathcal{V}_{\text{ach}}} \text{PV}_{i,v}. \tag{17}$$

Within each region $r \in \mathcal{R}$, we then take region-specific quartiles of ACHV_TOTAL and assign each student to one of four achievement levels:

$$\text{ach\_level}(i) \in \{\text{high}, \text{medium}, \text{low}, \text{poor}\}. \tag{18}$$

This yields a balanced representation of students at different overall achievement levels within each region, while avoiding domination by a few high-performing systems.

Psychological risk. To construct a binary psychological-risk indicator, we use the normalised psychological-symptoms index (e.g., PSYCHSYM; see Appendix K.1). Within each region, we define high-risk students as those whose z-score on PSYCHSYM lies in the top quartile, and low-risk students as all others:

$$\text{risk}(i) = \begin{cases} \text{high}, & z_{i,\text{PSYCHSYM}} \geq Q_{0.75}(r), \\ \text{low}, & \text{otherwise}, \end{cases} \tag{19}$$

where $Q_{0.75}(r)$ is the 75th percentile of PSYCHSYM z-scores in region r.

Cohort definition. Combining achievement level, gender, and psychological risk yields $4 \times 2 \times 2 = 16$ mutually exclusive cohorts per region:

$$\text{cohort}(i) = \big(\text{ach\_level}(i),\ \text{gender}(i),\ \text{risk}(i)\big). \tag{20}$$

For each region r, construct c, and cohort k, we compute the unweighted mean

$$\bar{z}_{r,c,k} = \frac{1}{|S_{r,k}|} \sum_{i \in S_{r,k}} z_{i,c}, \tag{21}$$

where $S_{r,k}$ is the set of students in region r and cohort k. The 16-dimensional vector $(\bar{z}_{r,c,k})_{k=1}^{16}$ is used as the real-student reference pattern for construct c in region r.

K.3 PISA-based Shadow Survey and Aggregation

The PISA-side shadow survey largely mirrors the CEPS design described in Appendix J.1, but is tailored to PISA constructs.

Item selection and translation. For each selected construct (Appendix K.1), we manually pick one to three representative items from the PISA student questionnaire that load strongly on that construct according to OECD documentation. Items are translated into Chinese while keeping the original response wording and Likert-type scales as close as possible. Negatively oriented items are reverse-coded so that higher numeric scores consistently indicate more of the underlying construct (e.g., higher anxiety or stronger self-efficacy).

Agent prompting and coding. On the agent side, we reuse the immersive role-playing prompt template in Figure 2. For each region and cohort, we sample a fixed number of personas, instantiate them as DeepSeek-V3.2-based student agents, and ask them to complete the PISA shadow survey. Free-text responses are post-processed with a simple rule-based mapper that aligns each answer to one of the discrete PISA response categories, which are then mapped to the official numeric codes used in the PISA scales.

Cohort-level aggregation and metrics. For each region, construct, and cohort, we compute the mean coded response across all sampled agents, yielding an agent-side vector of 16 cohort means that is directly comparable to the PISA vector $(\bar{z}_{r,c,k})_{k=1}^{16}$ defined in Appendix K.2. We then compute Pearson and Spearman correlations between the human and agent vectors exactly as in the CEPS setting (Appendix J.2), and report their distributions across regions and constructs in § 5.3.

K.4 PISA 2022 Constructs and Questionnaire Scales

Table 6 lists all PISA 2022 student questionnaire scales used in our cross-regional analysis, together with the latent families that we refer to in the main text (math engagement/efficacy, curiosity/growth, classroom climate/belonging, well-being, and workload/work–home balance). Throughout the paper we use the PISA short codes (e.g., MATHEFF, CURIOAGR) to denote these scales; this table provides the full mapping for readers who wish to inspect the underlying constructs in more detail.

Table 6: PISA 2022 constructs used in the agent–human alignment analysis. Each construct is measured as an OECD-provided index in the student questionnaire; the same construct name is used for human data (e.g., `MATHEFF`) and its agent counterpart (prefixed by `K_`, e.g., `K_MATHEFF`). Brief descriptions follow official PISA 2022 variable definitions.

Category	Construct	Variable name(s)	Brief description
Workload / practice	EXERPRAC	Human: `EXERPRAC`; Agent: `K_EXERPRAC`	Exercise or practice a sport before or after school
	STUDYHMW	Human: `STUDYHMW`; Agent: `K_STUDYHMW`	Studying for school or homework before or after school
	WORKHOME	Human: `WORKHOME`; Agent: `K_WORKHOME`	Working in household/take care of family members before or after school
	WORKPAY	Human: `WORKPAY`; Agent: `K_WORKPAY`	Working for pay before or after school
Math engagement & efficacy	MATHPREF	Human: `MATHPREF`; Agent: `K_MATHPREF`	Preference of Math over other core subjects
	MATHEASE	Human: `MATHEASE`; Agent: `K_MATHEASE`	Perception of Mathematics as easier than other subjects
	MATHMOT	Human: `MATHMOT`; Agent: `K_MATHMOT`	Motivation to do well in mathematics
	MATHEFF	Human: `MATHEFF`; Agent: `K_MATHEFF`	Mathematics self-efficacy: formal and applied mathematics - response options reversed in 2022 (WLE)
	MATHEF21	Human: `MATHEF21`; Agent: `K_MATHEF21`	Mathematics self-efficacy: mathematical reasoning and 21st century skills (WLE)
	MATHPERS	Human: `MATHPERS`; Agent: `K_MATHPERS`	Effort and Persistence in Mathematics (WLE)
	ANXMAT	Human: `ANXMAT`; Agent: `K_ANXMAT`	Mathematics Anxiety (WLE)
Classroom exposure	EXPOFA	Human: `EXPOFA`; Agent: `K_EXPOFA`	Exposure to Formal and Applied Mathematics Tasks (WLE)
	EXPO21ST	Human: `EXPO21ST`; Agent: `K_EXPO21ST`	Exposure to Mathematical Reasoning and 21st century mathematics tasks (WLE)
Classroom climate & belonging	RELATST	Human: `RELATST`; Agent: `K_RELATST`	Quality of student-teacher relationships (WLE)
	BELONG	Human: `BELONG`; Agent: `K_BELONG`	Sense of belonging (WLE)
	BULLIED	Human: `BULLIED`; Agent: `K_BULLIED`	Being bullied (WLE)
	SOCCON	Human: `SOCCON`; Agent: `K_SOCCON`	Social Connections: Ease of Communication About Worries and Concerns (WLE)
Curiosity & growth	CURIOAGR	Human: `CURIOAGR`; Agent: `K_CURIOAGR`	Curiosity (agreement) (WLE)
	EMOCOAGR	Human: `EMOCOAGR`; Agent: `K_EMOCOAGR`	Emotional control (agreement) (WLE)
	GROSAGR	Human: `GROSAGR`; Agent: `K_GROSAGR`	Growth Mindset (WLE)
Creativity & self-efficacy	CREATEFF	Human: `CREATEFF`; Agent: `K_CREATEFF`	Creative self-efficacy (WLE)
	CREATOP	Human: `CREATOP`; Agent: `K_CREATOP`	Creativity and Openness to Intellect TBD (WLE)
Well-being / mental health	LIFESAT	Human: `LIFESAT`; Agent: `K_LIFESAT`	Students' Life Satisfaction across Domains (WLE)
	PSYCHSYM	Human: `PSYCHSYM`; Agent: `K_PSYCHSYM`	Psychosomatic Symptoms (WLE)

L Additional Evaluation Results

L.1 Intrinsic Evaluation Results for TAD-PG Persona Collections

Corpus-level summary. Table 7 reports corpus-level intrinsic metrics for a sample of 151,426 HACHIMI personas used to answer RQ1. No persona triggered any hard schema error, and only a very small fraction (0.06%) received soft theoretical warnings.

Table 7: Summary of intrinsic metrics on 151,426 HACHIMI personas.

Metric	Value	Notes
# personas	151,426	merged corpus size
Error rate	0.0000	fraction with ≥ 1 hard error
Warning rate	0.0006	fraction with ≥ 1 warning
Distinct-1	0.4044	over values/creativity/psychology
Distinct-2	0.8296	over values/creativity/psychology
Near-duplicate pairs	0	SimHash, Hamming $\leq T$
SimHash Hamming (mean)	27.48	raw Hamming distance
SimHash Hamming (std)	4.10	across sampled pairs

Distributional alignment with quotas. To quantify how well the empirical distributions follow the scheduler-specified quotas, we compute KL divergence between the target and empirical distributions for each key categorical variable. On this corpus, the KL values are effectively zero for gender and academic level, and close to zero for grade (Table 8), indicating that the stratified quota mechanism realizes the intended marginal distributions up to negligible sampling noise.

Table 8: KL divergence between target quotas and empirical distributions.

Variable	KL Divergence
Grade	0.0001
Gender	0.0000
Academic level	0.0000

Semantic diversity and redundancy. Distinct-n statistics in Table 7 indicate substantial lexical and phrasal diversity across the three long-text components. The SimHash-based near-duplicate detector did not find any pairs below the Hamming threshold, and the average Hamming distance is roughly 27.5 bits with a standard deviation of 4.1, suggesting that profiles occupy a wide region of the semantic space rather than clustering around a few templates. Paragraph-length summaries for each component (values, creativity, mental health) further show tight control around target ranges with negligible proportions of overly short or long texts (see Figure 5).

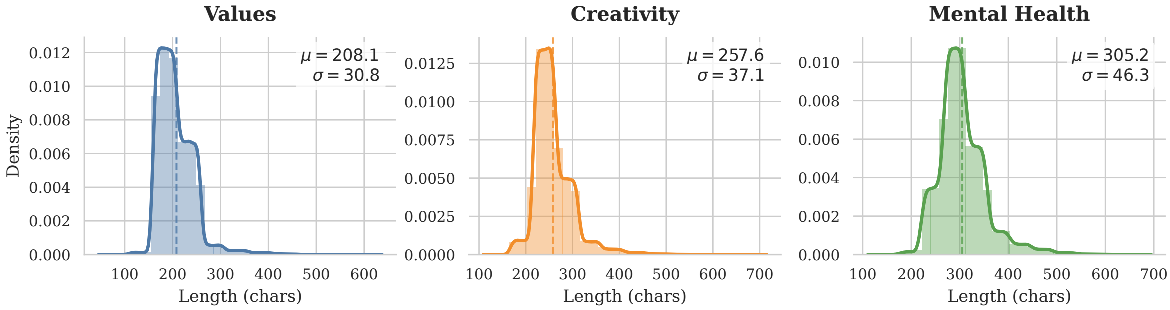

Figure 5: Distribution of paragraph lengths (in characters) across the three long-text components. The solid curves represent kernel density estimates (KDE), and the dashed lines indicate the means. The distributions show that the generation process exercises tight control over output lengths, centered around the target specifications.

Anchor alignment with academic level. Finally, we summarize the text-derived alignment scores between academic achievement and the three key components (values, creativity, mental health). For each persona, level words are mapped to numeric polarity scores and averaged within each component, then correlated with the four-level academic-achievement anchor. As shown in Table 9, all three correlations are positive and substantial, indicating that textual descriptions of high-achieving students tend to emphasize adaptive values, stronger creative profiles, and more favorable mental-health status in line with the TAD-PG design.

Table 9: Text-derived alignment between academic level and component scores.

Component	Pearson r
Values	0.7671
Creativity	0.8997
Mental health	0.8380

L.2 Additional Cross-regional Results on PISA 2022

To complement the summary in Section 5.3, this appendix provides additional detail on the cross-regional consistency analysis based on PISA 2022. We report results at the level of individual constructs (Table 6) and region–construct pairs, and include per-construct plots to allow readers to inspect where the agent-based personas align more or less closely with human group profiles.

Overall distribution across constructs and families. Within each region, we computed Pearson r and Spearman ρ between human and agent group means for every construct listed in Table 6, using the same 16 gender $\times$ achievement $\times$ risk groups as in the CEPS analysis. Figure 4 in the main text already summarizes these values as distributions by region. In the appendix, Figure 6 further breaks these distributions down by latent family, showing that math engagement/efficacy and curiosity/growth constructs systematically occupy the upper part of the correlation spectrum, whereas classroom climate/belonging constructs are more dispersed and well-being and workload constructs cluster near zero or negative values. Across regions, math and curiosity-related scales not only have the highest median correlations but also the smallest interquartile ranges, indicating that the strong alignment observed for `MATHEFF` and `CURIOAGR` is representative of a broader family-level pattern rather than being driven by a single favorable scale.

Regional variation at the construct level. Figure 7 presents a region $\times$ construct heatmap of Pearson correlations, making the construct-level regional differences discussed in Section 5.3 more explicit. Math engagement and efficacy constructs (`MATHEFF`, `MATHEASE`, `MATHEF21`, `MATHPERS`, `MATHPREF`) form a consistently high-correlation block across all five regions, with no region showing systematic degradation relative to the others. Curiosity/growth constructs (`CURIOAGR`, `GROSAGR`, `CREATOP`) are positively aligned everywhere but show slightly lower correlations in East Asia and Southern Europe than in Latin America and the Middle East, in line with the regional pattern described in the main text. By contrast, mental-health and well-being scales (`PSYCHSYM`, `LIFESAT`) tend to hover around $r \approx 0$ in every region, and

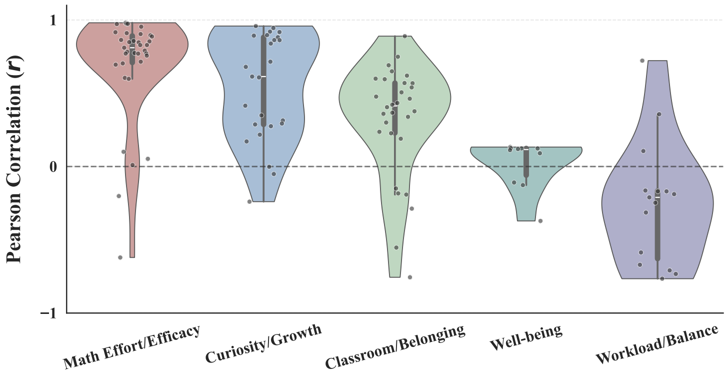

Figure 6: Distribution of Pearson correlations between human and agent group means on PISA 2022, summarized by construct family (each violin aggregates all constructs within a family across the five regions; individual construct points are jittered).

workload/work–home balance scales (WORKHOME, STUDYHMW, EXERPRAC) often fall into the negative range, indicating that agents systematically under- or over-estimate load for certain human groups. Some classroom-exposure variables, such as EXPO21ST, even flip sign between Southern Europe and other regions, illustrating that the agent-based personas can reproduce regional rank orders for academic constructs while remaining less reliable for specific aspects of classroom experience and well-being.

Finally, Table 10 reports the full set of Pearson and Spearman correlations for all region–construct pairs used in the analysis. Taken together with the family- and region-level visualizations, these detailed statistics reinforce the main conclusion in Section 5.3: agent-based personas reproduce a stable cross-regional structure for academic and curiosity-related constructs on PISA 2022, whereas well-being, workload, and certain classroom-exposure constructs exhibit weaker and more region-dependent alignment.

M Baseline Comparison Results

M.1 Baseline Definition and Evaluation Protocol

Baseline. We define a strong yet simple baseline that uses the same backbone LLM but generates each student persona in a *single pass* (i.e., one-shot holistic profiling) without multi-agent factorization, neuro-symbolic validation, quota scheduling, or semantic deduplication.

Protocol parity. Unless otherwise noted, all evaluations follow the exact same pipeline and hyper-parameters as HACHIMI: (i) the same schema parser and intrinsic evaluator (§I), (ii) the same cohort construction, sampling sizes, and correlation metrics for CEPS (§J) and PISA (§K), and (iii) the same student-agent instantiation and prompting template (Figure 2).

M.2 Intrinsic Comparison on 10,000 Personas

We report intrinsic metrics on a 10,000-persona subset generated by the baseline and compare them with a matched 10,000-persona subset from HACHIMI, using the same offline evaluator described in Appendix I.

Intrinsic comparison with a one-shot baseline. Table 11 compares HACHIMI with a matched one-shot single-model baseline on 10,000 personas, evaluated by the same offline intrinsic evaluator and thresholds as in Appendix I. HACHIMI exhibits substantially stronger structural reliability: the baseline yields a non-trivial hard-error rate (12.03%) and a high warning rate (25.33%), while HACHIMI incurs no hard errors and only a small fraction of soft warnings (0.82%). This suggests that single-pass holistic generation is prone to format deviations, missing theory-mandated slots, and cross-field inconsistencies, whereas the Propose–Validate–Revise loop effectively enforces the theory-anchored schema.

Diversity and redundancy. HACHIMI also shows higher lexical diversity (Distinct-1/2: 0.3285/0.7893 vs. 0.2328/0.4589). Consistently, SimHash indicates stronger redundancy in the baseline: 157 near-duplicate pairs are detected, and the distribution of pairwise Hamming distances shifts lower (mean 21.64

Table 10: Cross-regional consistency of Pearson correlations (r) between human and agent group means on PISA 2022. Constructs are grouped by family. **Bold** indicates strong alignment ($r \geq 0.80$); gray indicates negative alignment. Descriptions follow official PISA 2022 variable definitions.

Construct	Description	East Asia	S. Europe	Lat. Am.	Mid. East	W. Europe
Math Effort & Efficacy						
MATHEFF	Mathematics self-efficacy: formal and applied mathematics - response options reversed in 2022 (WLE)	**0.95**	**0.98**	**0.97**	**0.97**	**0.97**
MATHEASE	Perception of Mathematics as easier than other subjects	**0.83**	**0.81**	0.72	**0.91**	**0.85**
MATHEF21	Mathematics self-efficacy: mathematical reasoning and 21st century skills (WLE)	0.77	0.77	**0.90**	**0.89**	0.78
MATHPREF	Preference of Math over other core subjects	**0.83**	**0.92**	0.60	**0.90**	**0.85**
MATHPERS	Effort and Persistence in Mathematics (WLE)	**0.85**	0.70	0.76	**0.86**	0.79
ANXMAT	Mathematics Anxiety (WLE)	0.70	0.77	**0.86**	0.77	0.60
MATHMOT	Motivation to do well in mathematics	0.10	0.05	-0.20	-0.62	0.01
Curiosity & Growth						
CURIOAGR	Curiosity (agreement) (WLE)	**0.92**	**0.96**	**0.84**	**0.90**	**0.92**
EMOCOAGR	Emotional control (agreement) (WLE)	0.17	0.31	0.22	0.29	0.29
GROSAGR	Growth Mindset (WLE)	0.35	0.61	**0.88**	**0.89**	**0.86**
CREATEFF	Creative self-efficacy (WLE)	-0.05	0.42	-0.24	-0.00	0.28
CREATOP	Creativity and Openness to Intellect TBD (WLE)	0.61	**0.86**	0.72	0.68	**0.95**
Classroom & Belonging						
RELATST	Quality of student-teacher relationships (WLE)	0.54	0.60	0.48	0.62	0.75
BELONG	Sense of belonging (WLE)	0.38	0.37	0.60	0.30	0.24
BULLIED	Being bullied (WLE)	0.43	0.57	0.42		0.41
EXPOFA	Exposure to Formal and Applied Mathematics Tasks (WLE)	0.34	0.19	0.57	0.69	**0.89**
EXPO21ST	Exposure to Mathematical Reasoning and 21st century mathematics tasks (WLE)	0.23	-0.76	0.51	0.65	0.46
SOCCON	Social Connections: Ease of Communication About Worries and Concerns (WLE)	-0.18	-0.29	-0.15	-0.55	-0.19
Well-being						
PSYCHSYM	Psychosomatic Symptoms (WLE)	0.09	0.13	0.11	0.12	0.12
LIFESAT	Students' Life Satisfaction across Domains (WLE)	-0.11	0.12	-0.13	-0.37	0.12
Workload & Balance						
STUDYHMW	Studying for school or homework before or after school	0.72	0.11	-0.19	-0.31	0.36
EXERPRAC	Exercise or practice a sport before or after school	-0.25	-0.16	-0.21	-0.17	-0.17
WORKHOME	Working in household/take care of family members before or after school	-0.67	-0.71	-0.59	-0.77	-0.73

Table 11: Summary of intrinsic metrics for HACHIMI personas vs. the matched one-shot single-model baseline.

Metric	HACHIMI	Baseline	Notes
# personas	10,000	10,000	merged corpus size
Error rate	0.0000	0.1203	fraction with ≥ 1 hard error
Warning rate	0.0082	0.2533	fraction with ≥ 1 warning
Distinct-1	0.3285	0.2328	over values/creativity/psychology
Distinct-2	0.7893	0.4589	over values/creativity/psychology
Near-duplicate pairs	0	157	SimHash, Hamming $\leq T$
SimHash Hamming (mean)	29.86	21.64	raw Hamming distance
SimHash Hamming (std)	4.02	3.65	across sampled pairs

vs. 29.86), implying tighter semantic clustering and template reuse. Overall, modular generation with validation and deduplication yields personas that are both more schema-faithful and more diverse under the same sampling budget.

M.3 Full CEPS Results for the Baseline

Table 12 reports item- and construct-level consistency on CEPS Grade 8 by correlating the 16-dimensional cohort-mean vectors between humans and the one-shot baseline agents, using Pearson r and Spearman ρ (Appendix J.2). We additionally report Δr and $\Delta \rho$, computed as (HACHIMI $-$ Baseline), to quantify the relative advantage of our multi-agent persona generation pipeline.

Overall, HACHIMI exhibits consistently stronger human–agent alignment than the one-shot baseline, with improvements on nearly all targets in both rank-based and linear agreement. The gains are most pronounced on socially grounded and relational signals: teacher attention improves from ρ=0.7647 and r=0.7202 with $\Delta\rho = +0.1324$ and $\Delta r = +0.1390$; parent–child relationship quality shows large gains (father–child: $\Delta\rho = +0.1877$, $\Delta r = +0.2085$; mother–child: $\Delta\rho = +0.3309$, $\Delta r = +0.2659$). Notably, help-seeking when in trouble improves substantially ($\Delta\rho = +0.5362$, $\Delta r = +0.6410$), indicating that HACHIMI better preserves cohort-level patterns for behavioral and support-seeking tendencies.

Importantly, HACHIMI also strengthens alignment on key education-facing outcomes: educational aspiration and parental achievement expectations improve with $\Delta\rho = +0.2266$/+0.1848 and $\Delta r = +0.0856$/+0.1128, respectively, suggesting that the multi-agent workflow more faithfully captures population-level ordering and magnitude of academic-expectation constructs. Beyond these, HACHIMI yields broad improvements on perceived subject difficulty and psychosocial constructs, including mathematics difficulty ($\Delta\rho = +0.0940$, $\Delta r = +0.1180$), Chinese difficulty ($\Delta\rho = +0.1574$, $\Delta r = +0.1355$), depressive symptoms ($\Delta\rho = +0.0842$, $\Delta r = +0.0388$), and misbehavior frequency ($\Delta\rho = +0.0414$, $\Delta r = +0.0683$).

Meanwhile, parental strictness remains negatively correlated for both methods (Baseline: $\rho = -0.3324$, $r = -0.2733$; $\Delta\rho = -0.1823$, $\Delta r = -0.2295$), highlighting a challenging construct where static personas may not recover the same cohort gradients as the survey. Taken together, these results indicate that HACHIMI more reliably reproduces CEPS cohort-level structure than the one-shot baseline, especially for constructs tied to classroom experience, relationships, and observable behaviors.

M.4 Full PISA 2022 Results for the Baseline

Table 13 reports region-wise consistency on PISA 2022 by correlating the cohort-mean vectors between humans and the one-shot baseline agents, using Pearson r across the same five regions as in Appendix K. We additionally report per-region Δ, computed as (HACHIMI $-$ Baseline), to quantify the relative advantage of our multi-agent persona generation pipeline.

Overall, HACHIMI exhibits consistently stronger human–agent alignment than the one-shot baseline on most education-facing attitudes and classroom relationship constructs, with improvements that are broadly stable across regions. The gains are most pronounced on math self-beliefs and growth-oriented constructs: perceived ease of mathematics (`MATHEASE`) improves substantially from baseline r=0.45–0.63 with consistently large gains (Δ=+0.27–+0.29) across *all* regions; mathematics self-efficacy (`MATHEFF`) is already strong under the baseline (r=0.82–0.85) and is further strengthened by HACHIMI (Δ=+0.13–+0.14). Similarly, 21st-century math efficacy (`MATHEF21`) improves from r=0.62–0.77 with Δ=+0.08–+0.15, preference for mathematics (`MATHPREF`) improves from r=0.47–0.74 with Δ=+0.11–+0.18, and persistence on difficult math tasks (`MATHPERS`) improves from r=0.62–0.74 with Δ=+0.08–+0.17. Beyond math attitudes, HACHIMI yields broad improvements on growth and creativity-related self-perceptions: growth mindset (`GROSAGR`) gains are substantial in multiple regions (e.g., S. Europe Δ_{SE}=+0.28, W. Europe Δ_{WE}=+0.22), and creative self-efficacy (`CREATEFF`) improves consistently across all regions, mitigating negative or near-zero baseline correlations (e.g., East Asia: r=−0.15 with Δ_{EA}=+0.10; Lat. Am.: r=−0.38 with Δ_{LA}=+0.14; Mid. East: r=−0.26 with Δ_{ME}=+0.26; W. Europe: $r\approx0$ with Δ_{WE}=+0.29). Importantly, HACHIMI also strengthens alignment on classroom relational signals: student–teacher relationship quality (`RELATST`) improves from baseline r=0.33–0.46 with consistently positive gains (Δ=+0.13–+0.29), suggesting that the multi-agent workflow better preserves cohort-level patterns tied to classroom experience and interpersonal dynamics.

Meanwhile, several constructs remain challenging and exhibit limited or mixed improvements. Motivation to do well in mathematics (`MATHMOT`) is unstable under the baseline (near-zero or negative in some regions, e.g., Lat. Am. r=−0.28, Mid. East r=−0.63), and HACHIMI yields only marginal gains in most regions (Δ=+0.01–+0.08) and even a slight decrease in W. Europe (Δ_{WE}=−0.02). Exposure-related classroom constructs are also heterogeneous: while `EXPOFA` improves strongly in Lat. Am. (Δ_{LA}=+0.19), `EXPO21ST` remains particularly difficult, including a strongly negative baseline in S. Europe (r=−0.79) with only a small improvement (Δ_{SE}=+0.03), and slight regressions in East Asia and W. Europe (Δ_{EA}=−0.02, Δ_{WE}=−0.03). Notably, social connectedness (`SOCCON`) stays negative or near-zero under both methods and often regresses (e.g., Δ_{LA}=−0.12, Δ_{ME}=−0.29, Δ_{WE}=−0.19),

Table 12: Full CEPS consistency results for the baseline. We report Pearson r and Spearman ρ between human and baseline agent cohort means (16 cohorts). We additionally report Δr and $\Delta\rho$ computed as (HACHIMI − Baseline) on the same targets.

Target	Label / Description	Spearman ρ	Pearson r	$\Delta\rho$	Δr	p_ρ	Human / Agent mean range
Construct_Teacher_Attention_Avg	Teacher attention (avg)	0.7647	0.7202	+0.1324	+0.1390	1.54×10^{-5}	2.6-3.0 / 2.1-3.2
Construct_Depression	Depressive symptoms (sum)	0.2276	0.2859	+0.0842	+0.0388	0.0678	16.8-32.3 / 18.1-26.5
Construct_Parental_Strictness	Parental strictness (avg)	-0.3324	-0.2733	−0.1823	−0.2295	0.2085	2.7-2.9 / 2.1-2.3
Construct_Prosocial	Prosocial behaviour (avg)	0.4651	0.4684	+0.1673	+0.0935	0.0225	3.4-3.8 / 3.2-4.5
Construct_Misbehavior	Misbehaviour frequency (avg)	0.4615	0.3555	+0.0414	+0.0683	0.0236	1.5-2.3 / 1.2-1.8
Construct_School_Bonding	School bonding (avg)	0.1676	0.1807	+0.1853	+0.0906	0.3163	2.6-3.2 / 3.2-4.0
w2b18	Educational aspiration	0.7469	0.7798	+0.2266	+0.0856	2.77×10^{-8}	6.4-7.9 / 6.5-8.2
w2a27	Parental achievement expectations	0.7168	0.8493	+0.1848	+0.1128	1.13×10^{-4}	1.7-2.5 / 1.4-3.3
w2b21	Future confidence	0.5481	0.5667	+0.0773	+0.1070	0.0047	2.7-3.4 / 2.5-4.0
w2a29	Parental-expectation pressure	0.4421	0.5244	+0.0068	+0.0244	0.0073	2.6-3.3 / 2.0-3.0
w2a22	Father–child relationship quality	0.2004	0.2434	+0.1877	+0.2085	0.2583	2.2-2.7 / 2.9-3.0
w2a23	Mother–child relationship quality	0.3244	0.4599	+0.3309	+0.2659	0.4035	2.5-2.9 / 2.9-3.0
w2b02	Perceived difficulty in mathematics	0.7119	0.7380	+0.0940	+0.1180	1.34×10^{-4}	1.8-3.2 / 1.2-3.2
w2b03	Perceived difficulty in Chinese	0.2102	0.2475	+0.1574	+0.1355	0.6845	2.4-3.1 / 2.1-3.1
w2b04	Perceived difficulty in English	0.6794	0.7314	+0.1235	+0.1139	3.72×10^{-4}	1.7-3.0 / 1.2-3.1
w2c04	Overall self-rated health	0.1588	0.1752	+0.0093	+0.0370	0.3331	3.4-4.1 / 3.5-4.8
w2d13	Help-seeking when in trouble	-0.3091	-0.4001	+0.5362	+0.6410	0.2441	2.1-3.2 / 1.6-2.0

highlighting a boundary where static personas struggle to recover cohort gradients for latent or socially embedded well-being signals. Consistently, well-being and time-use variables show weak alignment and/or regression: psychosomatic symptoms (PSYCHSYM) remain close to zero (r=0.09–0.13) with negligible changes, life satisfaction (LIFESAT) shows mixed deltas including notable decreases in Mid. East ($\Delta_{\mathrm{ME}}=-0.23$) and Lat. Am. ($\Delta_{\mathrm{LA}}=-0.11$), and workload/balance measures (EXERPRAC, WORKHOME) consistently regress across regions (e.g., Δ down to -0.29 and -0.32, respectively). Taken together, these results indicate that HACHIMI more reliably reproduces PISA cohort-level structure than the one-shot baseline, especially for math attitudes, growth/creativity self-beliefs, and classroom relationship constructs, while social connectedness, well-being, and time-allocation patterns remain challenging for persona-based inference at population scale.

N Use of AI Assistants

GPT-5 was used to polish the appendix language, focusing on grammar and phrasing. All outputs were reviewed and revised by the authors. No AI tools used for scientific content or experiments.

Table 13: Baseline alignment on PISA 2022: Pearson correlations (r) between human and baseline agent cohort means across regions, with per-region Δ indicating HACHIMI$-$baseline. Bold indicates $r \geq 0.80$; gray indicates negative alignment. $\Delta > 0$ is blue and $\Delta < 0$ is red.

Construct	Description (official)	East Asia	Δ_{EA}	S. Europe	Δ_{SE}	Lat. Am.	Δ_{LA}	Mid. East	Δ_{ME}	W. Europe	Δ_{WE}
Math Effort & Efficacy											
MATHEFF	Mathematics self-efficacy: formal and applied mathematics - response options reversed in 2022 (WLE)	**0.82**	+0.13	**0.85**	+0.13	**0.84**	+0.13	**0.84**	+0.13	**0.83**	+0.14
MATHEASE	Perception of Mathematics as easier than other subjects	0.55	+0.28	0.52	+0.29	0.45	+0.27	0.63	+0.28	0.57	+0.28
MATHEF21	Mathematics self-efficacy: mathematical reasoning and 21st century skills (WLE)	0.62	+0.15	0.69	+0.08	0.76	+0.14	0.77	+0.12	0.67	+0.11
MATHPREF	Preference of Math over other core subjects	0.72	+0.11	0.74	+0.18	0.47	+0.13	0.72	+0.18	0.69	+0.16
MATHPERS	Effort and Persistence in Mathematics (WLE)	0.74	+0.11	0.62	+0.08	0.65	+0.11	0.70	+0.16	0.62	+0.17
ANXMAT	Mathematics Anxiety (WLE)	0.62	+0.08	0.71	+0.06	**0.80**	+0.06	0.69	+0.08	0.59	+0.01
MATHMOT	Motivation to do well in mathematics	0.05	+0.05	0.01	+0.04	-0.28	+0.08	-0.63	+0.01	0.03	-0.02
Curiosity & Growth											
CURIOAGR	Curiosity (agreement) (WLE)	**0.81**	+0.11	**0.94**	+0.02	**0.81**	+0.03	**0.89**	+0.01	**0.89**	+0.03
GROSAGR	Growth Mindset (WLE)	0.29	+0.06	0.33	+0.28	0.75	+0.13	0.73	+0.16	0.64	+0.22
CREATOP	Creativity and Openness to Intellect TBD (WLE)	0.37	+0.24	**0.86**	+0.00	0.70	+0.02	0.65	+0.03	**0.89**	+0.06
CREATEFF	Creative self-efficacy (WLE)	-0.15	+0.10	0.15	+0.27	-0.38	+0.14	-0.26	+0.26	-0.01	+0.29
EMOCOAGR	Emotional control (agreement) (WLE)	0.15	+0.02	0.31	+0.00	0.20	+0.02	0.27	+0.02	0.35	-0.06
Classroom & Belonging											
EXPOFA	Exposure to Formal and Applied Mathematics Tasks (WLE)	0.37	-0.03	0.17	+0.02	0.38	+0.19	0.66	+0.03	**0.84**	+0.05
RELATST	Quality of student-teacher relationships (WLE)	0.41	+0.13	0.41	+0.19	0.33	+0.15	0.46	+0.16	0.46	+0.29
EXPO21ST	Exposure to Mathematical Reasoning and 21st century mathematics tasks (WLE)	0.25	-0.02	-0.79	+0.03	0.40	+0.11	0.61	+0.04	0.49	-0.03
BULLIED	Being bullied (WLE)	0.36	+0.07	0.54	+0.03	0.42	+0.00	0.37	-0.01	0.36	+0.05
BELONG	Sense of belonging (WLE)	0.36	+0.02	0.39	-0.02	0.54	+0.06	0.25	+0.05	0.24	+0.00
SOCCON	Social Connections: Ease of Communication About Worries and Concerns (WLE)	-0.18	+0.00	-0.22	-0.07	-0.03	-0.12	-0.26	-0.29	0.00	-0.19
Well-being											
PSYCHSYM	Psychosomatic Symptoms (WLE)	0.09	+0.00	0.10	+0.03	0.11	+0.00	0.13	-0.01	0.11	+0.01
LIFESAT	Students' Life Satisfaction across Domains (WLE)	-0.07	-0.04	0.10	+0.02	-0.02	-0.11	-0.14	-0.23	0.02	+0.10
Workload & Balance											
STUDYHMW	Studying for school or homework before or after school	0.68	+0.04	0.10	+0.01	-0.20	+0.01	-0.32	+0.01	0.36	+0.00
EXERPRAC	Exercise or practice a sport before or after school	-0.15	-0.10	-0.10	-0.06	0.08	-0.29	0.07	-0.24	0.03	-0.20
WORKHOME	Working in household/take care of family members before or after school	-0.35	-0.32	-0.42	-0.29	-0.45	-0.14	-0.56	-0.21	-0.48	-0.25

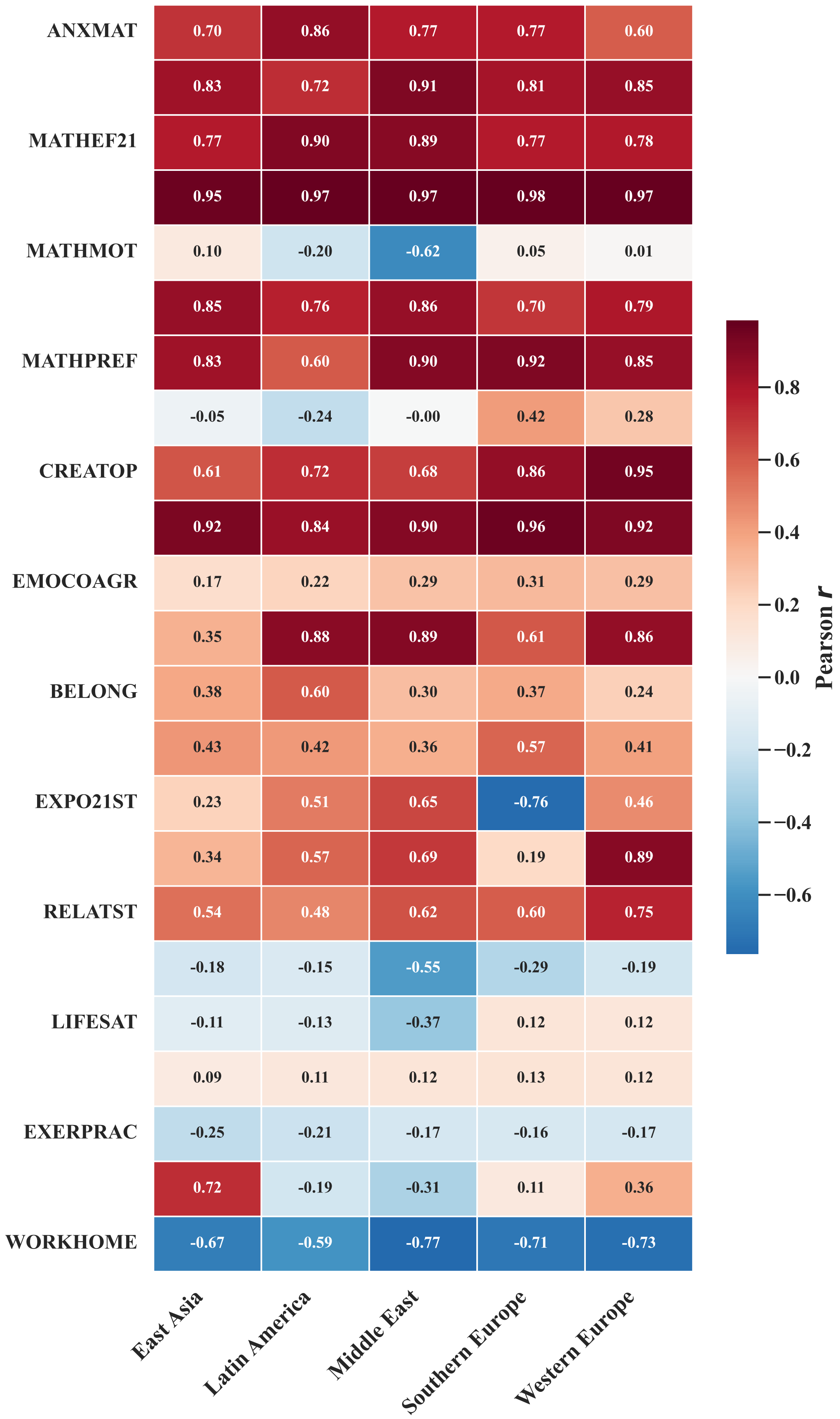

Figure 7: Heatmap of Pearson correlations between human and agent group means on PISA 2022. Constructs are ordered by latent family, highlighting the high-consistency math/curiosity block and the lower-consistency well-being and workload block.