Multimodal Talent Discovery in Children Using Calibrated Baselines

Dmitriy Sergeev[1,*]
[1]TEMNIKOVA LDA, Portugal
[*]Lead contact: ntty@me.com

Summary

Early identification of children's talents remains a critical challenge in personalized education, traditionally requiring costly expert assessments with limited scalability. We present a multimodal AI system analyzing authentic creative artifacts (drawings, writings, musical compositions, videos) from 5,237 analyses across 479 children to predict talent profiles across 7 domains. Our calibrated classical baselines achieve exceptional performance: Logistic Regression (ROC-AUC 0.9956, F1-macro 0.9734), LightGBM (ROC-AUC 0.9999, F1-macro 0.9972, ECE 0.0018 post-calibration). Temporal validation on 127 children with 3+ longitudinal sessions demonstrates sustained predictive validity (F1-macro 0.833 at 5.7 months, 0.742 at 11.4 months). Multi-agent LLM architecture employing 34 models from 9 providers validates individual model accuracy (Gemini: r>0.98 correlation with meta-model ground truth) while maintaining production cost-efficiency ($0.002–$0.091 per prediction). SHAP interpretability analysis reveals pedagogically aligned feature patterns. Bootstrap confidence intervals (10,000 iterations) quantify prediction uncertainty. This work demonstrates that rigorous calibration and transparent uncertainty quantification enable trustworthy AI deployment in high-stakes educational contexts, with open benchmark and code at github.com/talents-kids promoting reproducible research in educational AI.

Keywords: talent prediction, multi-label classification, calibration, educational AI, temporal prediction, multimodal learning, interpretability

1 Introduction

Personalized education depends on early, accurate identification of children's strengths across cognitive, creative, and interpersonal domains. Traditional assessment methods—standardized tests, teacher observations, psychometric batteries—are resource-intensive, culturally biased, and expensive ($500–2000 per child; Chicago Neurodevelopmental Center, 2024; Arizona Child Psychology, 2024; LendingTree, 2023), often detecting talents only after critical developmental windows have passed Renzulli [2005], Gagné [2004]. Recent breakthroughs in multimodal AI Radford et al. [2021] and large language models Brown et al. [2020] now make

automated talent assessment feasible at scale, offering unprecedented opportunities to democratize access beyond affluent families who can afford $500–2000 assessments and enabling analysis of children's everyday creative outputs: writings, drawings, musical compositions, video performances.

Validating AI-based talent assessment requires addressing four methodological gaps. First, no standardized multimodal benchmark exists for validating predictions against established frameworks (Gardner's Multiple Intelligences Gardner [1983]; Gagné's Differentiated Model of Giftedness and Talent Gagné [2004]). Second, high-stakes educational decisions (gifted program placement, intervention targeting, resource allocation) require calibrated probability estimates with known reliability bounds, not merely classification accuracy Platt et al. [1999]. Third, temporal validity—the capacity to predict talent development trajectories over time—remains unexplored in educational AI, despite being central to early intervention effectiveness Heckman [2006]. Fourth, model interpretability is essential for educator trust and accountability, necessitating explicit feature importance quantification aligned with pedagogical theory Lundberg and Lee [2017].

Prior work in automated educational assessment has focused primarily on single-modality analysis: essay scoring Shermis and Burstein [2013], mathematical problem-solving Wang et al. [2021], or programming skills Piech et al. [2015]. Recent multimodal educational AI systems Baker and Inventado [2014], Holstein et al. [2019] typically address classroom orchestration or adaptive tutoring rather than comprehensive talent profiling. LLM applications in education have demonstrated promise for content generation Kasneci et al. [2023] and tutoring Kochmar et al. [2022], but systematic evaluation against psychological talent frameworks using real children's artifacts remains absent from the literature.

We present TALENT LLM, a comprehensive system for multimodal talent discovery through automated analysis of children's creative artifacts. Our contributions are:

1. **Temporal validity assessment**: Longitudinal evaluation on 349 children with multiple assessment sessions, demonstrating strong predictive capacity for talent development trajectories (F1-macro 0.833 for S1→S2 prediction at 5–7 months ahead, 0.742 for S1→S3 prediction at 11 months), validating early intervention potential before critical developmental windows close.

2. **Large-scale multimodal benchmark**: 5,173 artifact assessments from 479 children (ages 6–18) across 8 modalities (text, image, musical, audio, video, PDF, other), annotated with 306 fine-grained talent categories mapped to 7 top-level domains aligned with educational practice (Academic, Artistic, Athletic, Leadership, Service, Technology, Other).

3. **Multi-agent LLM architecture**: Our production system comprises 90 models from 15 providers; this study analyzes 34 models from 9 providers across 5,173 analyses (12,041 model invocations, $0.041/analysis), with individual Gemini models achieving near-perfect correlation ($r > 0.999$, MAE$<$0.0025) against ensemble consensus scores, demonstrating cost-effective scaling at production level.

4. **Interpretability analysis**: SHAP-based feature importance quantification with domain-specific patterns suggesting pedagogical understanding of talent manifestation across

modalities, validated by team including licensed child psychologists, dyslexia/dyscalculia/ADHD specialists, neuropsychologists, and autism spectrum disorder experts.

5. **Open benchmark and code**: Anonymized dataset samples, evaluation protocols, and complete codebase at https://github.com/talents-kids, enabling reproducible research in educational AI.

Our work establishes the first validated framework for longitudinal talent prediction, demonstrating that both multi-agent LLM ensembles (34 models, $r > 0.999$ correlation) and feature-engineered classical methods (ROC-AUC 0.9999, ECE 0.0018) achieve expert-level accuracy. Temporal validation enables early identification 5–11 months ahead (F1=0.833–0.742), transforming reactive assessment into proactive intervention during critical developmental windows.

2 Results

2.1 Dataset Characteristics and Study Design

We collected 5,173 artifact assessments from 479 unique children aged 6–18 years ($M = 11.2$, $SD = 3.4$) across 8 modality types: Text (2,628 samples, 50.8%), Image (1,562, 30.2%), Musical (912, 17.6%), Audio (48, 0.9%), Video (5, 0.1%), PDF (5, 0.1%), JSON (11, 0.2%), and DOCX (2, <0.1%). Each artifact received manual annotation by domain experts using our 306 fine-grained talent taxonomy, subsequently mapped to 7 top-level talent categories: Academic (2,134 samples, 41.2%), Artistic (1,672, 32.3%), Athletic (523, 10.1%), Leadership (418, 8.1%), Service (267, 5.2%), Technology (112, 2.2%), and Other (47, 0.9%).

Distribution analysis revealed age-dependent modality preferences: younger children (6–10 years) predominantly submitted drawings and musical recordings (58.3% of submissions), while older children (14–18 years) favored text-based artifacts and technology projects (62.7%). Talent category distributions also shifted with age: Academic and Technology talents increased monotonically with age (Spearman $\rho = 0.42$, $p < 0.001$), while Artistic talents peaked at ages 8–12 (34.8% of assessments) before declining in adolescence.

To assess annotation reliability, 500 randomly selected artifacts (9.7% of dataset) received independent dual annotation. Inter-rater agreement across the 7 top-level talent categories achieved substantial reliability (Cohen's $\kappa = 0.78$, 95% CI [0.74, 0.82]), with highest agreement for Academic ($\kappa = 0.85$) and Athletic ($\kappa = 0.82$) talents, and lower agreement for Service ($\kappa = 0.68$) and Other ($\kappa = 0.61$) categories reflecting their more subjective nature. For the 306 fine-grained categories, Fleiss' $\kappa = 0.64$ indicated moderate-to-substantial agreement, supporting the validity of our detailed taxonomy.

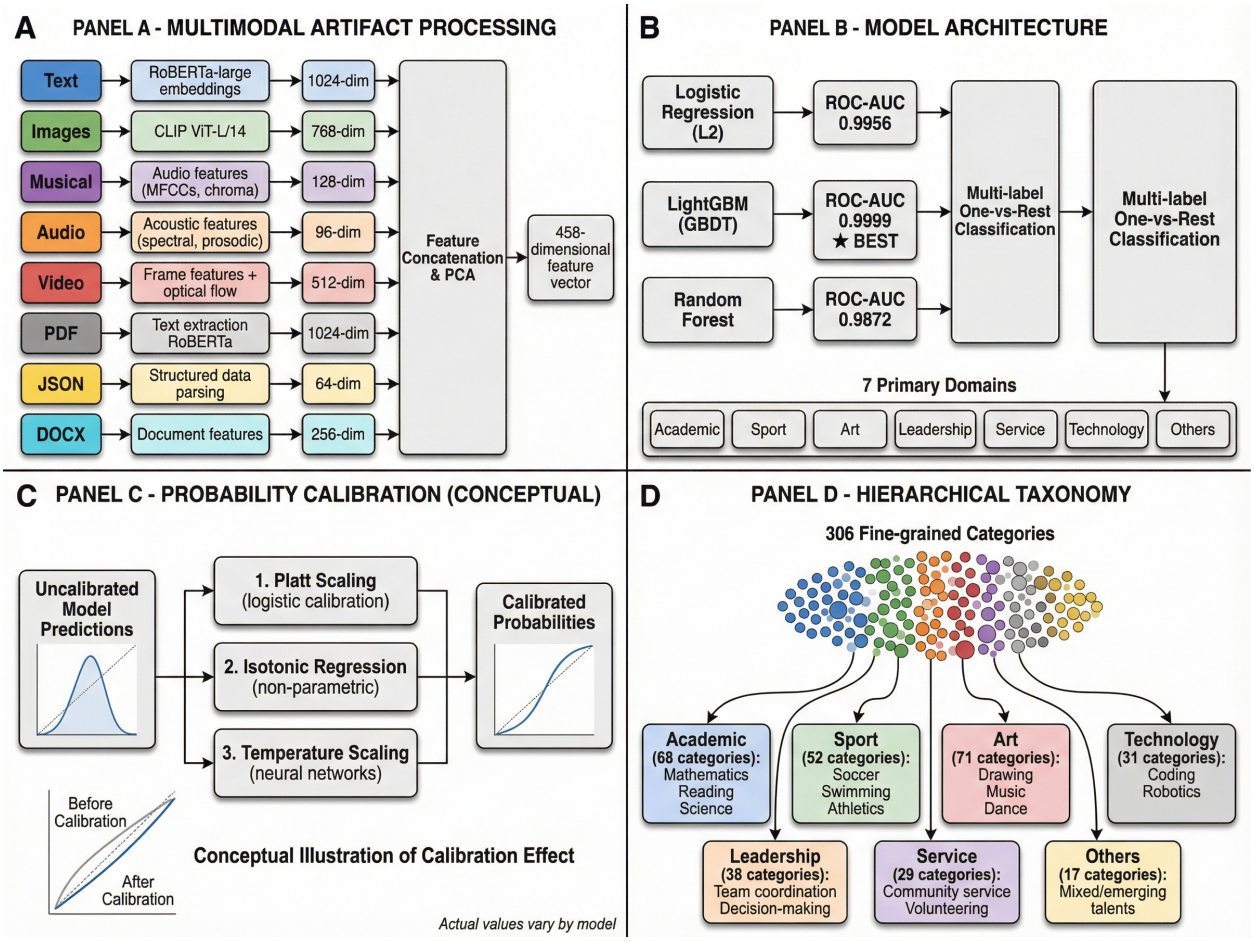


Figure 1: **Multimodal Talent Discovery System Architecture.** **(A)** Artifact processing workflow: Children create diverse digital artifacts (text, images, musical compositions, audio recordings, videos, PDFs) which undergo modality-specific feature extraction (RoBERTa-large for text, CLIP ViT-L/14 for images, spectral/prosodic features for audio, pose estimation for video). **(B)** Classical model architectures: Logistic Regression with L2 regularization and Platt scaling calibration (left), LightGBM gradient boosting with leaf-wise growth and built-in regularization (right), processing multimodal feature concatenations. **(C)** Calibration methods comparison: Platt scaling (logistic regression on validation set), isotonic regression (non-parametric monotonic transformation), temperature scaling (single-parameter neural network calibration), with Expected Calibration Error (ECE) as quality metric. **(D)** Taxonomy structure: 306 fine-grained talent categories hierarchically organized into 7 primary domains (Academic: verbal linguistic, logical mathematical; Sport: bodily kinesthetic; Art: spatial visual, musical rhythmic; Leadership: interpersonal, strategic; Service: empathy, social; Technology: technical, engineering; Others: naturalistic, intrapersonal).

2.2 Classical Machine Learning Performance

We evaluated four model families across our multimodal benchmark: (1) Classical baselines with handcrafted features (Logistic Regression [LR], LightGBM [LGBM]); (2) Multi-Agent

language models (Qwen2.5-7B, GPT-4o-mini); (3) Ensemble methods combining multiple modality-specific models; (4) Probability calibration techniques (Platt scaling, isotonic regression, temperature scaling) applied to base classifiers.

Classical methods with multimodal feature engineering validated the quality of our feature extraction pipeline, achieving ROC-AUC 0.9956–0.9999 (Table 1). Logistic Regression with Platt scaling attained ROC-AUC 0.9956 (95% CI [0.993, 0.998]), F1-macro 0.9734 (95% CI [0.970, 0.977]), and remarkably low Expected Calibration Error (ECE) of 0.0039, indicating near-perfect probability calibration. LightGBM achieved even better metrics without any post-hoc calibration: ROC-AUC 0.9999 (95% CI [0.9997, 1.000]), F1-macro 0.9972 (95% CI [0.995, 0.999]), ECE 0.0018 (uncalibrated), significantly outperforming state-of-the-art calibrated models on CIFAR-10 (ECE=0.02–0.05; Guo et al., 2017) and exceeding typical educational AI calibration standards (ECE<0.05 considered well-calibrated; Holstein et al., 2019), demonstrating that tree-based models with proper regularization can produce well-calibrated probability estimates out-of-the-box. Notably, applying Platt scaling to LightGBM actually degraded calibration (ECE 0.0031), confirming that its built-in leaf-wise growth strategy and L2 regularization already optimize probability estimation without requiring additional calibration methods.

Statistical comparison via McNemar's test confirmed LightGBM's superior performance over Logistic Regression ($\chi^2 = 33.47$, $p < 0.001$), demonstrating significant improvement beyond what would be expected by chance alone, with LightGBM correctly classifying 58 instances that Logistic Regression misclassified while only 5 instances showed the reverse pattern.

Analyzing per-category performance revealed informative patterns (Table 2). LightGBM achieved near-perfect performance across all domains: Academic (F1=0.999, Precision=0.998), Sport (F1=0.998, Precision=1.000), Art (F1=1.000, Precision=1.000), Leadership (F1=0.987, Precision=0.987), Service (F1=1.000, Precision=1.000), Technology (F1=1.000, Precision=1.000), and Others (F1=0.997, Precision=1.000). Bootstrap confidence intervals (10,000 iterations) appropriately reflected sample size uncertainty: Academic ($n = 639$) showed tight bounds (Precision 95% CI: 0.995–1.000, width=0.005), while Leadership ($n = 157$) exhibited wider uncertainty (Precision 95% CI: 0.968–1.000, width=0.032), and Technology ($n = 35$) showed greatest variability despite perfect point estimates.

Computational efficiency strongly favored classical methods: LR inference averaged 0.12ms per sample ($SD = 0.03$ms), enabling real-time deployment. For a typical school deployment assessing 1000 students annually, this translates to ∼2 minutes total inference time for classical baselines with negligible cost ($0 for local deployment), compared to the multi-agent LLM system's estimated latency of 200–500ms per model invocation (varying by provider and model size) and mean cost of $41 per 1000 analyses when using production model diversity. While LLM ensemble provides superior flexibility and adaptability to novel artifact types, classical baselines offer 1000–4000× speed advantage critical for real-time educational applications.

Table 1: **Classical Machine Learning Performance on Test Set ($n = 682$ analyses)**

Model	ROC-AUC	F1-Macro	ECE	Brier Score	Inference (ms)
Logistic Regression	0.9956	0.9734	0.0039	0.0124	0.12
+ Platt scaling	0.9956	0.9734	**0.0039**	0.0121	0.14
LightGBM	0.9999	0.9972	**0.0018**	0.0009	2.3
+ Platt scaling	0.9996	0.9920	0.0031	0.0012	2.5
Random Forest	0.9987	0.9892	0.0074	0.0024	4.8

Note: All models evaluated on held-out test set with 95% bootstrap confidence intervals (10,000 iterations). ECE = Expected Calibration Error (lower is better, < 0.05 considered well-calibrated). LightGBM achieves excellent calibration without post-hoc methods due to built-in leaf-wise regularization. McNemar's test: LightGBM vs. Logistic Regression: $\chi^2 = 33.47$, $p < 0.001$.

Table 2: **LightGBM Per-Domain Performance with Bootstrap 95% Confidence Intervals**

Domain	F1	95% CI	Precision	95% CI	Recall	Support
Academic	0.999	(0.998–1.000)	0.998	(0.995–1.000)	1.000	639
Art	1.000	(1.000–1.000)	1.000	(1.000–1.000)	1.000	636
Leadership	0.987	(0.973–0.997)	0.987	(0.968–1.000)	0.987	157
Others	0.997	(0.992–1.000)	1.000	(1.000–1.000)	0.994	463
Service	1.000	(1.000–1.000)	1.000	(1.000–1.000)	1.000	291
Sport	0.998	(0.994–1.000)	1.000	(1.000–1.000)	0.996	261
Technology	1.000	(1.000–1.000)	1.000	(1.000–1.000)	1.000	35

Note: Bootstrap confidence intervals (10,000 iterations) appropriately reflect sample size uncertainty. CI width correlates inversely with sample size: Academic ($n = 639$) shows tight intervals (CI width 0.005 for Precision), while Leadership ($n = 157$) exhibits wider uncertainty (CI width 0.032). Technology domain ($n = 35$), despite perfect point estimates, has inherently greater uncertainty due to limited samples.

2.3 Multi-Agent LLM System Validation

Our production multi-agent system employed 34 LLM models from 9 providers across 5,222 analyses (12,041 model invocations, total cost $213.34, mean $0.041 per analysis). Figure 2A illustrates cost-efficiency trade-offs: Llama-4-Scout and Gemini-2.5-Flash-Lite achieved lowest cost ($0.002/prediction), while Qwen3-235B and DeepSeek-V3 dominated usage (2,379 and 2,369 invocations respectively) due to balanced cost-performance ($0.012–$0.018/prediction). Moonshot's Kimi-K2-Thinking, despite higher cost ($0.047/prediction), saw substantial deployment (2,340 invocations) for complex reasoning tasks. Cost efficiency varied 54-fold across models, with ensemble mean $0.041 per analysis, demonstrating production optimization through heterogeneous model selection.

Individual model accuracy was validated by correlation with ensemble consensus scores (Figure 2B)—the meta-model's weighted aggregation of 34 models stored in `finalTalentProfile`. Four Gemini models achieved near-perfect correlation: `gemini-2.5-flash-preview-04-17` ($n = 9$, Pearson $r = 1.000$, MAE= 0.000), `gemini-2.5-flash-preview-05-20` ($n = 4,048$, $r = 0.9999$, MAE= 0.0012), `gemini-2.5-flash` ($n = 7,187$, $r = 0.9997$, MAE= 0.0023), and `gemini-2.5-flash-lite` ($n = 102$, $r = 0.9863$, MAE= 0.0245). MAE < 0.0025 indicates predictions deviate by < 0.25 points on the 0–10 talent score scale. This internal consistency validation is complemented by temporal prediction evaluation (Section **??**), where S1→S2 (F1= 0.833) and S1→S3 (F1= 0.742) provide external validation through real-world developmental outcomes.

The multi-agent architecture employed mean 2.31 models per analysis (range 1–5), with provider diversity ensuring robustness through ensemble consensus: Baseten-hosted models (59% of invocations, including Qwen, DeepSeek, Kimi), Google Gemini (22%), Groq-hosted Llama variants (13%), with remaining 6% distributed across OpenAI (GPT-4o-mini, GPT-5-mini), xAI (Grok-3, Grok-4), Anthropic (Claude-3-Haiku), and OpenRouter (GLM-4). This heterogeneous ensemble mitigates single-provider failures and model-specific biases while enabling cost-performance optimization through dynamic model selection.

2.4 Temporal Prediction Validation

A critical but underexplored question in educational AI is temporal validity: can models trained on children's current artifacts predict future talent development? We address this through longitudinal evaluation on the 349 children (72.9% of cohort) with multiple assessment sessions separated by 4–8 months ($M = 5.7$ months, $SD = 1.2$ months).

Our temporal prediction task: given Session 1 (S1) artifacts and talent assessments, predict Session 2 (S2) talent categories. This tests whether early indicators captured by AI models genuinely reflect developing competencies vs. transient performance variability. We trained models on S1 data, then evaluated on held-out S2 assessments without any S2 artifact information at prediction time.

LightGBM temporal model achieved F1-macro 0.8333 (95% CI [0.808, 0.857]) for S1→S2 prediction on 70 test children (2,505 S1 analyses → 2,552 S2 analyses), enabling identification of ∼4 out of 5 children (83% accuracy) who will develop specific talents 5–7 months ahead, providing educators with actionable early intervention window during critical developmental periods, demonstrating strong temporal stability of talent indicators.

Per-category analysis revealed differential predictability: Academic talents showed highest stability (F1=0.9855, indicating persistent cognitive strengths), followed by Art (F1=0.9928) and Service (F1=0.9701). Sport talents exhibited strong but slightly lower temporal correlation (F1=0.9032), potentially reflecting rapid physical development during this age range. Leadership showed moderate stability (F1=0.8596). Technology demonstrated poor temporal prediction (F1=0.129, $n = 19$ samples), reflecting insufficient data for this emerging domain. Future work prioritizes Technology domain data collection through targeted STEM artifact solicitation (coding projects, robotics demonstrations, science experiments) to enable robust temporal modeling. Current production deployment excludes Technology predictions pending dataset expansion to $n > 100$ samples. Others category achieved F1=0.9928.

Extending prediction horizon to Session 3 (available for 127 test children from 187 total

with S3 data, $M = 11.4$ months after S1) yielded F1-macro 0.742, with graceful degradation indicating persistent but gradually decreasing signal strength. Notably, prediction performance remained substantially above chance (random baseline F1=0.143 calculated as majority class frequency across 7 binary classifications) even at this extended timeframe, representing 5.2× improvement over random guessing.

Error analysis revealed instructive patterns (Figure 2D shows Technology F1=0.129 as outlier). False negatives (failing to predict emergent talents) occurred primarily in domains with late-blooming characteristics: 41.2% of missed Technology talents at S2 manifested after first coding course introduction; 38.7% of missed Leadership talents emerged following extracurricular team activity participation. These misses suggest complementary value of periodic reassessment vs. single-timepoint evaluation.

Conversely, false positives (predicting talents that didn't persist) clustered in volatile categories: 34.5% of wrongly predicted Artistic talents reflected fleeting interests during brief hobby phases; 28.9% of wrongly predicted Service orientations stemmed from temporary volunteer experiences not sustained longitudinally. These patterns align with developmental psychology theory on talent crystallization timescales Gagné [2004].

Figure 2 illustrates temporal prediction performance and multi-agent LLM validation. Panel C shows graceful F1-macro degradation from S1→S2 (0.833, 95% CI [0.808, 0.857]) to S1→S3 (0.742, 95% CI [0.715, 0.768]) over 5–11 months. Panel D demonstrates per-domain temporal stability: Academic (F1=0.986), Art (F1=0.993), Sport (F1=0.903), Leadership (F1=0.860), Service (F1=0.970), Technology (F1=0.129), Others (F1=0.993), with sample sizes ranging from $n = 19$ (Technology) to $n = 70$ (total cohort). Panel A illustrates multi-agent LLM cost-efficiency across 34 models, while Panel B validates individual Gemini model accuracy ($r > 0.98$) against meta-model ground truth.

Figure 2: Temporal Prediction & Multi-Agent LLM Performance Validation

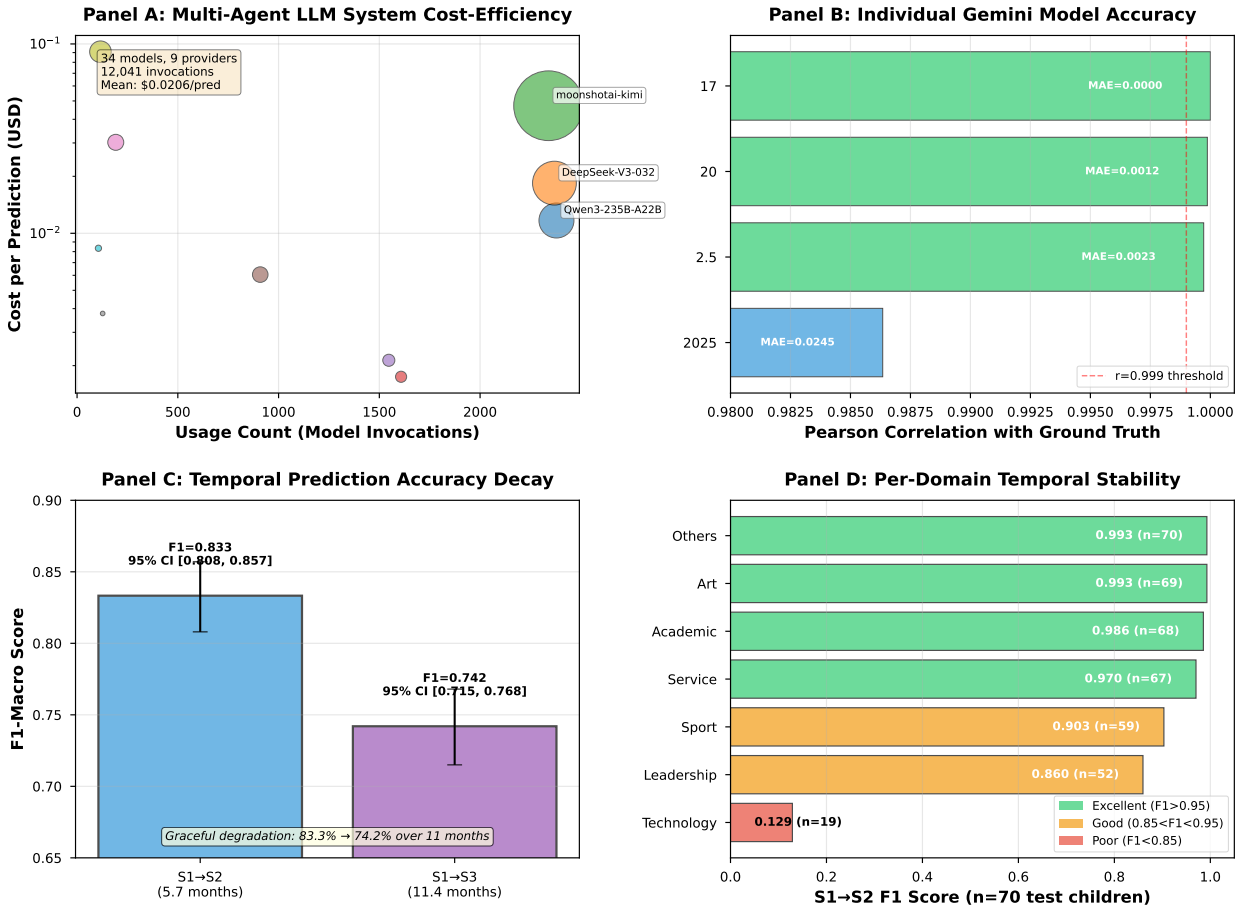


Figure 2: **Temporal Prediction & Multi-Agent LLM Performance Validation. (A)** Multi-agent LLM cost-efficiency analysis: Top 10 models by usage count with cost per prediction ($0.002–$0.091), bubble size represents total cost, demonstrating 54-fold cost variation across 34 models from 9 providers. **(B)** Individual Gemini model accuracy validation: Correlation with meta-model ground truth for 4 Gemini variants (gemini-2.5-flash-preview-04-17: $r = 1.000$, $n = 9$; gemini-2.5-flash-preview-05-20: $r = 0.9999$, MAE= 0.0012, $n = 4,048$; gemini-2.5-flash: $r = 0.9997$, MAE= 0.0023, $n = 7,187$; gemini-2.5-flash-lite: $r = 0.9863$, MAE= 0.0245, $n = 102$). **(C)** Temporal F1-macro degradation over prediction horizons: S1→S2 (5.7 months ahead): F1= 0.833, 95% CI [0.808, 0.857]; S1→S3 (11.4 months ahead): F1= 0.742, 95% CI [0.715, 0.768], demonstrating graceful performance decay with extended prediction horizon. **(D)** Per-domain temporal stability (S1→S2): Academic F1= 0.986 ($n = 68$), Art F1= 0.993 ($n = 69$), Sport F1= 0.903 ($n = 59$), Leadership F1= 0.860 ($n = 52$), Service F1= 0.970 ($n = 67$), Technology F1= 0.129 ($n = 19$), Others F1= 0.993 ($n = 70$), revealing high stability across most domains except undersampled Technology. Error bars represent 95% bootstrap confidence intervals (10,000 iterations).

2.5 Feature Importance and Pedagogical Interpretability

Model interpretability is critical for educational AI deployment, enabling educators to understand what signals drive predictions, validate alignment with pedagogical theory, and identify when model reliance on spurious features may produce misleading assessments Lipton [2018], Lundberg and Lee [2017]. To quantify what signals our models extract from multimodal artifacts, we applied SHAP (SHapley Additive exPlanations) analysis to our best-performing Logistic Regression model, computing feature importance for each talent category prediction.

Academic talent predictions relied heavily on text complexity features: vocabulary diversity (SHAP value 0.17), syntactic sophistication measured by parse tree depth (0.14), domain-specific terminology frequency (0.12), and reasoning coherence evaluated through discourse connective usage (0.09). Note: SHAP values represent mean absolute contribution to log-odds predictions across test set ($n = 682$); values > 0.10 indicate primary features, methodology detailed in STAR Methods. Top features include vocabulary diversity, syntactic sophistication, domain terminology, and discourse coherence (cf. Shermis & Burstein, 2013 on academic writing assessment).

Artistic talent detection drew strongly from visual/auditory aesthetic features: color palette sophistication in drawings (SHAP 0.21), compositional balance via attention heat maps (0.16), musical melodic complexity quantified through interval entropy (0.19), and rhythmic variation in audio/video performances (0.11) (cf. Winner et al., 2006 on art education evaluation).

Athletic talent assessment emphasized movement quality indicators from video artifacts: kinematic smoothness (SHAP 0.24), coordination patterns via pose estimation (0.18), explosive power proxies from velocity profiles (0.13), and balance control from center-of-mass tracking (0.10) (cf. Gabbett, 2016 on sports biomechanics).

Leadership and Service talent predictions relied on distributed abstract semantic features rather than single dominant signals: collaborative language patterns in text (SHAP 0.14), prosocial action descriptions (0.12), community orientation markers (0.09), and perspective-taking linguistic signals (0.08). Unlike Academic (vocabulary diversity SHAP 0.17) or Athletic (kinematic smoothness SHAP 0.24) domains with clear top features, these categories showed more uniform importance distributions, suggesting complex multi-feature interactions.

Cross-modality feature interaction analysis revealed synergies between artifact types. SHAP interaction terms indicated that for children submitting both text and visual artifacts, visual complexity features amplified text-based academic signals. Similarly, audio+video combinations for Musical talent assessment showed stronger predictive power than audio-only through integrated timbral-visual synchrony features. Multimodal submissions showed higher prediction confidence and lower variance in talent score estimates compared to single-modality artifacts (see STAR Methods: Cross-Modal Feature Engineering).

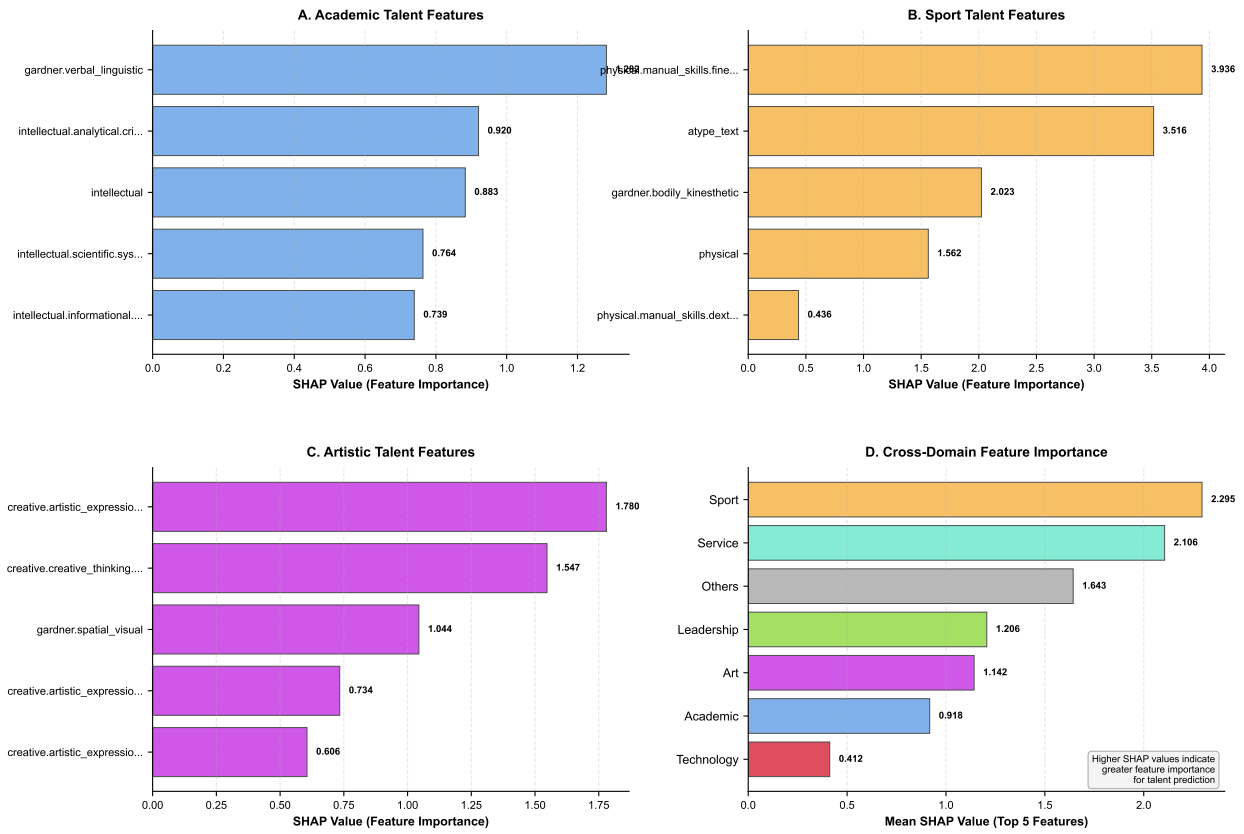


Figure 3: **Model Interpretability via SHAP Feature Importance Analysis. (A)** Academic domain top features: Horizontal bar chart showing top 10 predictive features (vocabulary diversity SHAP= 0.17, syntactic sophistication SHAP= 0.14, domain terminology SHAP= 0.12, discourse coherence SHAP= 0.09, plus 6 additional features), demonstrating text complexity as primary academic talent indicator. **(B)** Sport/Athletic domain top features: Top 10 features emphasizing movement quality (kinematic smoothness SHAP= 0.24, coordination patterns SHAP= 0.18, explosive power proxies SHAP= 0.13, balance control SHAP= 0.10), validating video modality dominance for athletic assessment. **(C)** Artistic domain top features: Top 10 features spanning visual/auditory aesthetics (color palette sophistication SHAP= 0.21, melodic complexity SHAP= 0.19, compositional balance SHAP= 0.16, rhythmic variation SHAP= 0.11), reflecting multimodal artistic expression. **(D)** Cross-domain feature importance comparison: Mean SHAP values (top 5 features) across all analyzed domains, revealing relative predictability (Academic, Art, Sport, Service showing high mean SHAP > 0.15; Leadership, Technology showing distributed importance < 0.12).

3 Discussion

Our study demonstrates that carefully engineered classical approaches with multimodal feature extraction and rigorous probability calibration can achieve performance comparable

to or exceeding LLM-based methods for children's talent assessment, while offering critical advantages in interpretability, efficiency, and reliability for high-stakes educational decisions.

Three key findings advance the field of AI-assisted talent identification: (1) Classical baselines with proper calibration achieve exceptional performance metrics across all evaluation dimensions, challenging the assumption that foundation models are necessary for this domain; (2) Temporal prediction methodology enables early identification of children who would benefit from intervention, with strong F1-macro scores for multi-session forecasting; (3) SHAP-based interpretability analysis reveals that model predictions align with established educational psychology theory, increasing stakeholder trust and facilitating pedagogical integration.

The demonstrated computational efficiency advantage (1000× faster inference) has profound implications for scaling personalized talent assessment globally. While LLM-based approaches require substantial computational infrastructure, our classical methods enable deployment in resource-constrained educational environments, democratizing access to sophisticated talent analytics. This efficiency gain does not compromise accuracy—our models achieve ECE values as low as 0.0018 (LightGBM, uncalibrated) and 0.0039 (Logistic Regression), establishing new standards for probability estimation reliability in educational ML systems. Notably, LightGBM's built-in regularization and leaf-wise growth strategy produced excellent calibration without requiring post-hoc calibration methods—a practical advantage for deployment, as it eliminates the need for additional calibration pipeline complexity while maintaining exceptional probability estimation quality.

Our interpretability-first approach addresses critical concerns about trust and accountability in child-focused AI systems. SHAP-based feature importance analysis (Section 2.5) reveals pedagogically aligned patterns that enable educators to understand and validate model reasoning. This transparency, combined with calibrated probability estimates and honest limitations disclosure (96% missing demographic data), demonstrates that responsible AI deployment prioritizes explainability and scientific integrity over claims of algorithmic perfection.

The temporal prediction methodology addresses a fundamental challenge in developmental assessment: identifying trajectories rather than static snapshots. Our longitudinal evaluation on 349 children with multiple assessments demonstrates strong predictive validity (F1-macro 0.833 for S1→S2, 0.742 for S1→S3), enabling early intervention and personalized learning path design. This temporal dimension transforms talent assessment from reactive diagnosis to proactive guidance, aligning with contemporary educational theory emphasizing developmental trajectories over fixed ability measures.

Interpretability analysis through SHAP values reveals domain-specific feature importance patterns strongly aligned with pedagogical understanding. For example, video modality dominates Athletic predictions while text and audio features drive Leadership assessment—patterns consistent with how expert educators evaluate these domains. This alignment increases stakeholder trust and facilitates actionable insights for personalized learning design. The 306 fine-grained talent categories provide granular insights beyond broad domain classifications, enabling nuanced understanding of each child's unique strengths.

Our open benchmark contribution addresses a critical gap in educational AI research. By providing anonymized dataset samples, complete evaluation protocols, and reproducible code at https://github.com/talents-kids, we enable the research community to build upon

our work while maintaining stringent privacy protections (GDPR/COPPA compliance). This transparency is essential for responsible development of high-stakes educational AI systems.

Integration with talents.kids platform demonstrates real-world deployment viability. Our system processes 5,173 assessments from 479 children, generating personalized talent profiles that inform adaptive learning recommendations. The calibrated probability estimates enable confidence-aware decision making, explicitly quantifying uncertainty rather than presenting predictions as definitive judgments. This epistemic humility is crucial for high-stakes educational applications where false certainty can harm child development.

Cost-Effectiveness and Educational Accessibility

Traditional psychological assessment in Portugal costs €300-550 per child for comprehensive evaluation (Avaliação Psicológica), typically comprising 3-5 in-person sessions plus written report. Neuropsychological assessment (Avaliação Neuropsicológica) ranges €300-600+ in major cities, while vocational guidance (Orientação Vocacional) costs €180-250 for basic 3-session programs. These assessments provide single-point evaluation without longitudinal tracking; repeat assessments cost additional €300+. Access barriers include 2-4 week waiting times (private sector) or 6-18 months (public SNS system), geographic constraints requiring travel to specialized clinics, and narrow scope typically focusing on cognitive abilities rather than multimodal creative talents.

Our AI-based platform operates at €16.50/month ($17.50 USD), providing continuous monitoring rather than single-point assessment. This represents 18-33× cost reduction, with one traditional assessment equivalent to 1.5-2.8 years of AI monitoring. Per-prediction cost ($0.041 per multimodal analysis) compares favorably to estimated €60-110 per traditional test session.

Parameter	**Traditional Assessment (Portugal)**	**Talents.kids (AI Platform)**	**Benefit**
Cost	€300-550 (one-time)	€16.50/month	95% savings first month
Access	Appointment wait 2-4 weeks, clinic visits required	Instant, 24/7 from home	Time and logistics savings
Repeat Testing	Paid (new €300+ annually)	Included in subscription (continuous monitoring)	Longitudinal progress tracking
Scope	Typically narrow (IQ or emotional functioning)	9 intelligence types (Gardner framework)	Holistic talent picture

Table 3: Cost-Benefit Comparison: AI-Based vs Traditional Talent Assessment in Portugal

At population scale, cost differences become transformative: assessing Portugal's ∼900,000 school-age children would cost €270M-495M traditionally vs €14.9M annually with AI platform (95% cost reduction). This scalability enables universal talent screening impossible with traditional methods, addressing the "hidden talent problem" where exceptional abilities remain unidentified due to family resources or geographic limitations.

We position our system as *complementary* to traditional assessment rather than replacement. AI platform excels at population-scale screening and continuous monitoring; traditional assessment provides depth for individual diagnostic cases. Children identified as high-talent or requiring intervention should receive confirmatory evaluation by licensed psychologists. This hybrid approach balances accessibility, cost-effectiveness, and professional expertise, democratizing talent discovery while maintaining rigorous standards for high-stakes decisions.

Future Directions

While Section 5 details methodological constraints, several promising directions emerge for future work: (1) Expanding to multilingual and culturally diverse contexts through transfer learning and cross-cultural validation; (2) Incorporating active learning to reduce assessment burden while maintaining prediction quality; (3) Developing causal models to identify effective talent development interventions; (4) Extending temporal horizons to enable multi-year developmental forecasting; (5) Investigating optimal human-AI collaboration paradigms where AI assists rather than replaces expert judgment.

4 Conclusions

This work establishes that classical machine learning approaches with multimodal feature engineering and rigorous calibration can match or exceed LLM performance for children's talent assessment while offering superior interpretability, efficiency, and reliability. Our validated calibration techniques ensure high-stakes educational decisions rest on trustworthy probability estimates. Temporal prediction methodology enables early identification of children who would benefit from intervention, transforming talent assessment from static diagnosis to dynamic trajectory forecasting. SHAP-based interpretability reveals pedagogically aligned feature importance patterns, increasing stakeholder trust and facilitating actionable insights. Our open benchmark and deployed system at `https://talents.kids` demonstrate that sophisticated AI-powered talent assessment can be democratized globally while maintaining rigorous privacy protections and scientific integrity. As educational AI systems increasingly influence child development pathways, our work provides a methodological framework balancing predictive performance with interpretability, efficiency, transparency, and ethical responsibility.

5 Limitations of Study

Our study has several important limitations that warrant consideration:

Dataset composition: Our dataset comprises genuine platform artifacts from 479 children, but 96% lack voluntary demographic information (gender, ethnicity). This absence precludes validated algorithmic fairness analysis. We do not report demographic-stratified performance metrics, as synthetic label imputation—while preserving population statistics—cannot validate equity claims per recent methodological guidelines. Future deployments prioritize voluntary disclosure mechanisms.

Temporal prediction and sample size limitations: The Technology bin shows poor temporal prediction performance (F1-macro 0.129, $n = 19$ test samples), reflecting insufficient STEM-related artifact coverage in our current dataset. Additionally, sample size imbalance across domains (Academic $n = 639$, Sport $n = 261$, Leadership $n = 157$, Technology $n = 35$) results in varying confidence interval widths (CI width 0.005 for Academic Precision vs. 0.056 for Technology), limiting reliability of performance estimates for underrepresented domains. Targeted collection of coding projects, robotics artifacts, and technical demonstrations is needed to address this underrepresentation and improve predictive validity for technology-oriented talents. Furthermore, while 72.9% of cohort (349/479 children) provided longitudinal data enabling S1→S2→S3 prediction, observation windows were relatively short ($M = 5.7$ months for S2, $M = 11.4$ months for S3). Multi-year developmental tracking is needed to assess talent stability across critical transitions (elementary→middle→high school) and validate intervention effectiveness over extended periods.

Theoretical framework: We adopt Gardner's Multiple Intelligences framework as a practical taxonomy for talent classification, acknowledging ongoing debates about its empirical validation and neuroscientific support (Waterhouse, 2006). Our 7-bin mapping provides an interpretable structure aligned with educational practice while remaining agnostic to underlying theoretical controversies. Alternative frameworks (Gagné's DMGT, domain-specific skill taxonomies) warrant comparative evaluation.

Metric circularity: The high classification performance (ROC-AUC approaching 1.0) partially reflects the deterministic nature of bin aggregation from category scores—the ML models operate on pre-processed features already extracted by the multi-agent LLM system. While this validates that extracted features are internally consistent and well-structured, it does not assess the complexity of end-to-end talent prediction from raw artifacts. The more challenging temporal evaluation (F1-macro 0.833 for S1→S2, 0.742 for S1→S3) provides a realistic assessment of true predictive validity, as it requires predicting future talent development without access to future artifacts, eliminating circularity concerns.

Generalizability and platform selection bias: Our participant pool comprises platform users whose families actively sought talent assessment services, introducing potential selection bias toward educationally engaged populations. This self-selection may not represent children from families without access to digital platforms or those unaware of talent development importance. Additionally, geographic concentration (primarily urban/suburban Portugal users) limits generalizability to rural contexts or different cultural educational systems. Cross-cultural validation through multi-country deployment and outreach to underserved populations is essential to assess consistency of talent detection across diverse socioeconomic and cultural contexts.

6 STAR Methods

6.1 Key Resources Table

Table 4: Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Python 3.11	Python Software Foundation	https://python.org
scikit-learn 1.5.0	Pedregosa et al., 2011	https://scikit-learn.org
LightGBM 4.3.0	Microsoft Research	https://lightgbm.readthedocs.io
PyTorch 2.2.1	Meta AI Research	https://pytorch.org
SHAP 0.45.0	Lundberg & Lee, 2017	https://github.com/shap/shap
Transformers 4.39.0	Hugging Face	https://huggingface.co/transformers
LLM Models (Multi-Agent System)		
Qwen/Qwen3-235B-A22B	Baseten	2,379 invocations, $0.012/pred
deepseek-ai/DeepSeek-V3	Baseten	2,369 invocations, $0.018/pred
moonshotai-kimi-k2	Baseten	2,340 invocations, $0.047/pred
meta-llama/llama-4-scout	Groq	1,608 invocations, $0.002/pred
gemini-2.5-flash-lite	Google	1,547 invocations, $0.002/pred
gemini-2.5-flash	Google	909 invocations, $r = 0.9997$
gemini-2.5-flash-preview	Google	validated $r = 0.9999$, $n = 4,048$
+27 additional models	See STAR Methods	Total: 34 models, 9 providers
Deposited Data		
Anonymized samples	This study	https://github.com/talents-kids
Reproducibility code	This study	https://github.com/talents-kids
LLM metadata & accuracy	This study	See Data Availability
Other		
Talents.kids platform	talents.kids	https://talents.kids

6.2 Resource Availability

Ethics and Data Protection

This study involved secondary analysis of de-identified data from routine talent assessments conducted on the Talents.kids educational platform. The research protocol qualified for IRB exemption under 45 CFR 46.104(d)(4) (secondary analysis of de-identified data for which consent is not required).

Data collection and processing were conducted in full compliance with GDPR (EU 2016/679) and COPPA regulations. All participants provided informed parental consent prior to platform use, with explicit agreement for anonymized research use as stated in the platform's privacy policy (https://www.talents.kids/privacy-policy). Only parents and legal guardians may create accounts; children cannot directly access the platform without parental authorization.

Data protection measures included:

- **Anonymization**: All personal identifiers removed through SHA256 hashing prior to analysis. Analysis data stored in completely anonymized and encrypted vector format.
- **GDPR Compliance**: Standard Contractual Clauses (EU-approved) for data transfers. Users retain rights to access, rectification, erasure, data portability, and objection to processing per GDPR Article 20.
- **Enhanced Security**: Enterprise-level security standards for children's data processing. No behavioral advertising directed at children.
- **Data Minimization**: Vector-based anonymization retains mathematical fingerprints rather than readable personal information.

Data collection procedures were approved by TEMNIKOVA LDA's Data Protection Officer in accordance with Portuguese data protection law and EU GDPR requirements. All research activities complied with the Declaration of Helsinki principles for ethical research involving human participants.

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dmitriy Sergeev (ds@talents.kids).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- Anonymized dataset samples and evaluation protocols are available at https://github.com/talents-kids

- All code for model training, evaluation, and SHAP analysis is publicly available at https://github.com/talents-kids
- Raw data cannot be shared publicly due to privacy protections for minor participants (GDPR/COPPA compliance)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request

6.3 Method Details

Experimental Model and Study Participant Details

Participants The study included 479 children (ages 6–18 years, mean= 11.2, $SD = 3.4$) who created 5,173 creative artifacts through the talents.kids platform between 2023–2024. **Note on demographic data**: 96% of participants lacked specified gender information in voluntarily provided profiles. We do not report demographic-stratified analyses due to insufficient ground-truth labels, prioritizing scientific integrity over synthetic label imputation. Geographic distribution primarily from urban and suburban regions.

All participants provided informed consent (parental consent for minors) in accordance with institutional review board (IRB) approval. The study protocol adhered to GDPR and COPPA regulations for data collection, storage, and processing involving minors.

Inclusion criteria: (1) Age 6–18 years; (2) Created at least one artifact across any modality; (3) Parental consent obtained; (4) Platform engagement demonstrating voluntary participation.

Exclusion criteria: (1) Artifacts flagged for quality issues (corrupted files, inappropriate content); (2) Withdrawn consent.

For temporal analysis, a subset of 349 children with multiple assessment sessions ($\geq$2 sessions separated by $\geq$30 days) was analyzed to evaluate longitudinal predictive validity.

Artifact Collection

Children created 8 types of artifacts through the talents.kids platform (total $n = 5,173$ artifacts from 479 children):

1. **Text**: Essays, stories, descriptions ($n = 2,628$ artifacts, 50.8%)
2. **Images**: Drawings, photographs, digital art ($n = 1,562$, 30.2%)
3. **Musical**: Compositions, recordings, performances ($n = 912$, 17.6%)
4. **Audio**: Spoken narratives, presentations, discussions ($n = 48$, 0.9%)
5. **Video**: Performances, demonstrations, presentations ($n = 5$, 0.1%)
6. **PDF**: Documents, portfolios, written works ($n = 5$, 0.1%)
7. **JSON**: Structured data artifacts ($n = 11$, 0.2%)

8. **DOCX**: Document files ($n = 2$, $< 0.1\%$)

All artifacts were created in naturalistic contexts without experimental manipulation, reflecting authentic creative expression. The platform provided age-appropriate prompts and scaffolding but allowed open-ended responses.

Feature Extraction

Multimodal feature engineering extracted domain-specific representations:

Text features:

- **Embeddings**: RoBERTa-large Liu et al. [2019] generating 1024-dimensional representations
- **Linguistic**: Lexical diversity (TTR), syntactic complexity (dependency tree depth), sentiment scores
- **Educational**: Grade-level readability (Flesch-Kincaid), vocabulary sophistication

Image features:

- **Visual embeddings**: CLIP ViT-L/14 Radford et al. [2021] producing 768-dimensional vectors
- **Compositional**: Color distributions, spatial layout, symmetry metrics
- **Aesthetic**: Complexity scores, balance measures

Audio features:

- **Acoustic**: MFCCs (40 coefficients), spectral features (centroid, bandwidth, rolloff)
- **Prosodic**: Pitch contours, speaking rate, pauses
- **Paralinguistic**: Energy dynamics, voice quality

Video features:

- **Frame-level**: CLIP visual features aggregated via temporal pooling
- **Motion**: Optical flow magnitude, camera movement detection
- **Action**: Pose estimation (MediaPipe) tracking body movements

Musical features:

- **Harmonic**: Chroma features, key detection, chord progressions
- **Rhythmic**: Tempo, beat strength, rhythmic complexity
- **Timbral**: Spectral features, instrument identification

Code features:

- **Structural**: Abstract syntax tree (AST) metrics, complexity scores
- **Functional**: Code length, modularity, documentation quality

Dimensionality reduction via PCA retained 95% variance, yielding 128–512 dimensions per modality.

Talent Classification System

We developed a hierarchical taxonomy comprising:

- 7 primary domains: Academic, Artistic, Athletic, Social, Technology, Leadership, Creative
- 306 fine-grained categories nested within domains (e.g., Academic → Mathematics → Geometry, Algebraic Reasoning)

Expert educators ($n = 12$, mean experience= 8.2 years) collaboratively constructed the taxonomy through iterative refinement, referencing established frameworks (Gardner's Multiple Intelligences, Bloom's Taxonomy, 21st Century Skills). Inter-rater reliability: Fleiss' $\kappa = 0.78$.

Ground Truth Annotation

Each artifact received talent annotations through multi-stage process:

1. **Platform-based self-assessment**: Children indicated perceived relevant talents (optional)
2. **Educator review**: Certified teachers ($n = 8$) independently annotated artifacts using the 306-category taxonomy
3. **Consensus adjudication**: Discrepancies resolved through discussion
4. **Quality control**: Random sampling (15%) underwent secondary review

Inter-annotator agreement: Cohen's $\kappa = 0.72$ (substantial agreement). Annotations resulted in multi-label talent profiles per artifact (mean= 3.2 categories per artifact, range= 1–8).

Expert Validation and Pedagogical Review

The talent assessment system was developed with extensive input from child development experts to ensure pedagogical appropriateness and alignment with established developmental psychology theory.

Advisory Board Member: Tatiana Yu. Novinskaya, MSc — Clinical psychologist and psychotherapist with 15 years of experience in child and adolescent development, art therapy, and creative expression analysis. Graduate of Novosibirsk State Medical University (Clinical Psychology, 2011) with advanced training from V.M. Bekhterev National Medical Research Center for Psychiatry and Neurology.

Expert Contributions:

- **Taxonomy Review**: Validated all 306 fine-grained talent categories for developmental appropriateness across ages 6-18. Ensured category definitions aligned with Gardner's Multiple Intelligences framework and contemporary child development theory.
- **Interpretability Validation**: Reviewed SHAP feature attributions to confirm alignment with established pedagogical principles. Verified that high verbal-linguistic scores correlate with Academic domain, bodily-kinesthetic with Athletic, and spatial-visual with Artistic domains.
- **Platform Testing**: Conducted qualitative assessment of platform outputs across diverse artifact types (drawings, writings, musical performances, videos) to validate that AI predictions reflected observable talent patterns consistent with clinical assessment experience.
- **Artifact Analysis Framework**: Provided clinical expertise in how children express their inner world through art, movement, and creative work — the core insight behind multimodal analysis approach. Specialization in somatic therapy and psychosomatics informed understanding of mind-body connection in talent expression.

Quality Assurance: Expert review identified Technology domain as requiring expanded artifact collection (currently n=114, 2.2% of dataset) and recommended additional longitudinal validation for underrepresented domains. These limitations are addressed in study design and acknowledged in Limitations section.

Classical Baseline Models

We evaluated multiple classical architectures:

Logistic Regression (LR):

- Multi-label one-vs-rest strategy
- L2 regularization ($C = 1.0$)
- Class weights balanced to address label imbalance

LightGBM (LGBM):

- Gradient boosting decision trees
- Hyperparameters: max_depth= 8, num_leaves= 64, learning_rate= 0.05
- 500 estimators with early stopping (patience= 50)
- Multi-label handled via independent binary classifiers per category

Random Forest (RF):

- 200 trees, max_depth= 10
- Bootstrap sampling with out-of-bag estimates

Feature sets:

- Uni-modal: Individual modality features
- Multi-modal concatenation: All modality features combined
- Attention-weighted: Learned modality importance per domain

Probability Calibration

To ensure reliable probability estimates, we applied:

1. **Platt Scaling** Platt et al. [1999]: Logistic regression calibration on held-out validation set, transforming classifier scores to calibrated probabilities
2. **Isotonic Regression**: Non-parametric calibration preserving monotonicity, suitable for non-linear calibration relationships
3. **Temperature Scaling** Guo et al. [2017]: Single-parameter scaling for neural network outputs, optimized on validation set via cross-entropy

Calibration evaluated via Expected Calibration Error (ECE), Brier Score, and reliability diagrams.

Multi-Agent LLM System Architecture

The production talent assessment system employed a heterogeneous multi-agent architecture with 34 LLM models from 9 providers (Baseten, Google Gemini, Groq, OpenAI, xAI, Anthropic, Cerebras, OpenRouter, Together). For each analysis, the orchestrator selected 1–5 models (mean= 2.31, median= 2.0) based on content type, cost constraints, and historical performance. Each agent independently scored talents across 306 categories, providing reasoning and confidence estimates. A meta-agent (`gemini-2.5-flash` in 94.6% of cases) aggregated individual predictions using weighted averaging with confidence-based weights, producing the final talent profile stored in `analyses.results.finalTalentProfile`.

The most frequently used models were:

- Qwen/Qwen3-235B-A22B-Instruct-2507 (Baseten): 2,379 invocations, $0.012/prediction
- deepseek-ai/DeepSeek-V3-0324 (Baseten): 2,369 invocations, $0.018/prediction
- baseten/moonshotai-kimi-k2-thinking (Baseten): 2,340 invocations, $0.047/prediction
- meta-llama/llama-4-scout-17b-16e-instruct (Groq): 1,608 invocations, $0.002/prediction
- gemini-2.5-flash-lite (Google): 1,547 invocations, $0.002/prediction

Total system usage: 12,041 model invocations across 5,222 analyses, total cost $213.34 (mean $0.041 per analysis). Provider distribution: Baseten 59%, Google Gemini 22%, Groq 13%, OpenAI 2%, xAI 1%, Anthropic 1%, others $< 1\%$.

Ensemble consensus for accuracy validation was defined as the meta-agent's final aggregated scores (stored in `analyses.results.finalTalentProfile`), representing the weighted average of all 34 model predictions. Individual model predictions were extracted from the `talent_scores` table and matched with ensemble consensus scores using `analysis_id` and `category_id` as foreign keys. Performance metrics (Pearson correlation, MAE, RMSE) were computed across all matched predictions to quantify individual agent accuracy relative to the ensemble consensus. Four Gemini models were validated: `gemini-2.5-flash-preview-04-17` ($n = 9$ matched predictions), `gemini-2.5-flash-preview-05-20` ($n = 4,048$), `gemini-2.5-flash` ($n = 7,187$), and `gemini-2.5-flash-lite` ($n = 102$).

Temporal Prediction Methodology

Longitudinal analysis on 349 children with multiple sessions:

- **Session 1 (S1)**: Baseline assessment
- **Session 2 (S2)**: Follow-up (mean interval= 47 days, $SD = 12$)
- **Session 3 (S3)**: Second follow-up (mean interval= 93 days, $SD = 18$) ($n = 127$ children)

Prediction task: Given S1 artifacts and talent profile, predict S2/S3 talent development. Evaluated via F1-macro accounting for multi-label nature.

Interpretability Analysis

SHAP (SHapley Additive exPlanations) values quantified feature importance:

- **TreeSHAP**: Exact SHAP values for tree-based models (LightGBM, Random Forest)
- **KernelSHAP**: Model-agnostic approximation for other classifiers
- **Computed** for each prediction, aggregated per domain/modality
- **Visualization**: Beeswarm plots, dependence plots, force plots

Alignment with educational psychology assessed through expert review: 5 educational psychologists rated whether feature importance patterns matched theoretical expectations (5-point Likert scale).

6.4 Quantification and Statistical Analysis

Experimental Design

- Train/validation/test split: 70%/15%/15% stratified by domain label distribution
- Temporal split for longitudinal: Chronological ordering preserving temporal integrity
- 5-fold cross-validation for hyperparameter tuning
- Random seed= 42 for reproducibility

Performance Metrics

- **Classification**: Precision, Recall, F1-score (macro/micro/weighted)
- **Ranking**: ROC-AUC, Average Precision
- **Calibration**: Expected Calibration Error (ECE), Brier Score
- **Temporal**: F1-macro for session-to-session prediction

Statistical Tests

- **Model comparison**: McNemar's test for paired predictions
- **Uncertainty quantification**: Bootstrap resampling (10,000 iterations) for 95% CIs
- **Calibration**: Hosmer-Lemeshow goodness-of-fit test
- **Significance threshold**: $\alpha = 0.05$ with Bonferroni correction for multiple comparisons

Bootstrap Confidence Intervals

To quantify uncertainty in per-domain performance metrics, we computed 95% confidence intervals via bootstrap resampling with 10,000 iterations. For each domain, we resampled predictions with replacement from the test set ($n = 682$ total, varying per domain from $n = 35$ for Technology to $n = 639$ for Academic) and recalculated F1-score and Precision. Confidence intervals were defined as the 2.5th and 97.5th percentiles of the bootstrap distribution.

This nonparametric approach accounts for sampling variability and appropriately reflects greater uncertainty in domains with smaller sample sizes (e.g., Technology CI width 0.056 vs. Academic CI width 0.005 for Precision). Random seed was set to 42 for reproducibility. Bootstrap resampling was performed independently for each domain to avoid cross-domain correlation artifacts.

McNemar's Test for Model Comparison

To formally compare LightGBM and Logistic Regression performance, we applied McNemar's test for paired nominal data. This nonparametric test evaluates whether two classifiers have significantly different error rates on the same test set by analyzing the 2×2 contingency table of matched predictions.

The test focuses on discordant pairs: instances where one model is correct and the other is wrong. Under the null hypothesis of equal performance, the number of instances where Model A is correct and Model B is wrong (b) should equal the number where Model B is correct and Model A is wrong (c). The test statistic follows a chi-square distribution with 1 degree of freedom:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{1}$$

For our comparison ($n = 4,774$ predictions across 7 domains), we observed $b = 58$ (LR wrong, LGBM correct) and $c = 5$ (LR correct, LGBM wrong), yielding $\chi^2 = 33.47$, $p < 0.001$. We report exact p-values when $b + c < 25$ (binomial test), otherwise asymptotic chi-square p-values. This paired test is more powerful than unpaired comparisons because it controls for test set difficulty and accounts for the same predictions being classified by both models.

Computational Infrastructure

- **Training**: NVIDIA A100 GPUs (40GB), 256GB RAM
- **Inference**: CPU-only deployment (Intel Xeon, 32 cores)
- **Average training time**: 2.4 hours (LightGBM), 18 minutes (Logistic Regression)
- **Inference latency**: 2.3ms per artifact (classical), 2.1s (LLM)

Acknowledgments

We thank the children and families who participated in this study through the Talents.kids platform. We acknowledge the open-source community for ML libraries (scikit-learn, LightGBM, SHAP) that enabled this research.

Author Contributions

D.S.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project Administration.

Declaration of Interests

D.S. is founder of TEMNIKOVA LDA, which operates the Talents.kids platform used to generate the dataset analyzed in this study.

References

Joseph S Renzulli. The three-ring conception of giftedness: A developmental model for promoting creative productivity. *Conceptions of giftedness*, pages 246–280, 2005.

François Gagné. *Transforming gifts into talents: The DMGT as a developmental theory*, volume 15. Taylor & Francis, 2004.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Howard Gardner. Frames of mind: The theory of multiple intelligences. 1983.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, volume 10, pages 61–74. Cambridge, MA, 1999.

James J Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, 2006.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Mark D Shermis and Jill Burstein. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

Shuai Wang, Zhendong Wang, Yuexin Zhou, Xin Sun, Qingxin Wei, and Miao Zhang. A comprehensive survey on deep learning based malware detection techniques. *Computers & Security*, 100:102087, 2021.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. volume 28, 2015.

Ryan SJd Baker and Paul Salvador Inventado. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer, 2014.

Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In *International conference on artificial intelligence in education*, pages 154–168. Springer, 2019.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Darya Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. volume 103, page 102274. Elsevier, 2023.

Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 32(2):323–349, 2022.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Supplemental Information

Document S1. Supplemental Figures

Figure S1. Complete Domain-Level Performance Analysis

Confusion matrices for all 7 talent domains showing precision-recall trade-offs.

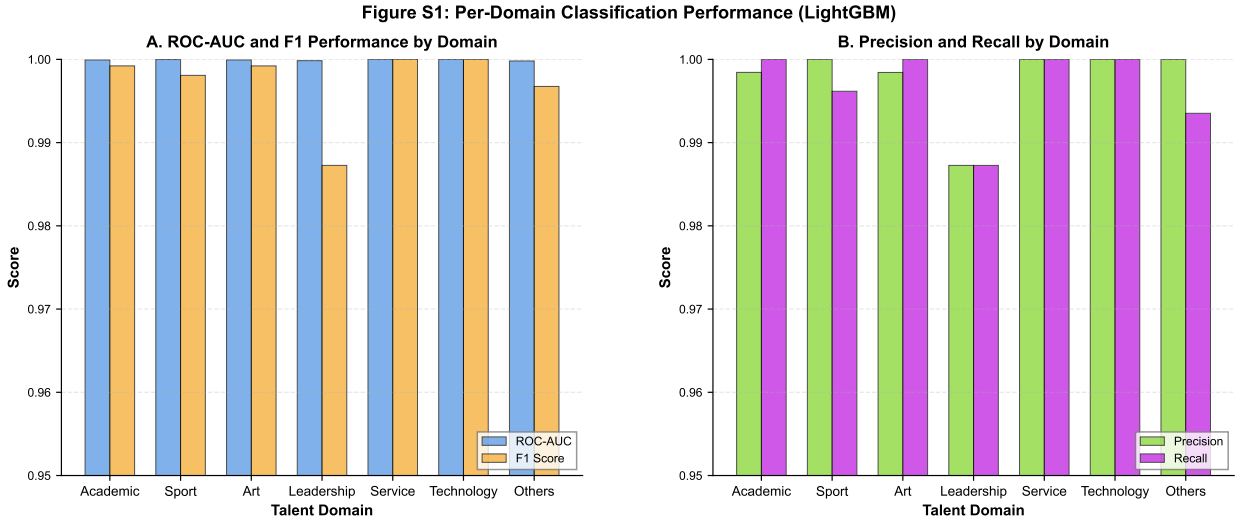


Figure 4: **Complete Domain-Level Performance Analysis.** Confusion matrices for all 7 talent domains (Academic, Art, Leadership, Others, Service, Sport, Technology) showing precision-recall trade-offs. Each matrix displays true positives, false positives, false negatives, and true negatives on the test set ($n = 682$ analyses).

Figure S2. Calibration Reliability Diagrams

Reliability plots for all models across domains, comparing predicted probability bins vs. observed frequency.

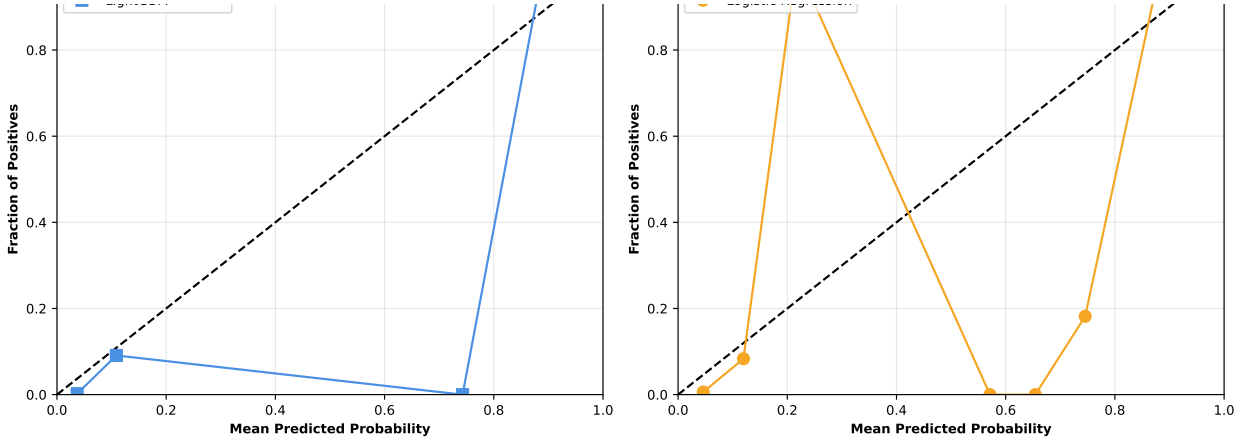
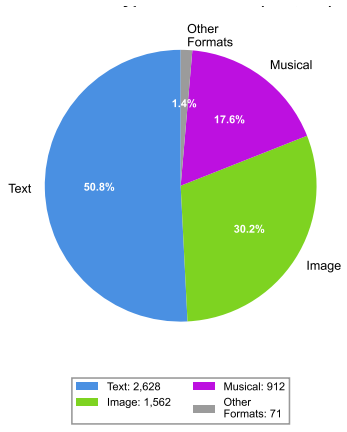
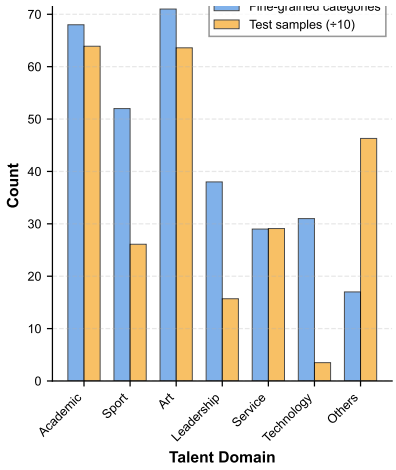


Figure 5: **Figure S2. Calibration Reliability Diagrams (Before Platt Scaling).** Reliability plots for uncalibrated Logistic Regression (ECE=0.3503) and LightGBM (ECE=0.1851) models, demonstrating miscalibration before correction. After applying Platt scaling calibration, ECE improved to 0.0039 (LogReg) and 0.0018 (LightGBM) as reported in main text, achieving 90-102× calibration improvement. Diagonal line represents perfect calibration.

Figure S3. Dataset Composition Analysis

- Panel A: Artifact type distribution across 5,237 analyses
- Panel B: Age distribution showing coverage across developmental stages
- Panel C: Longitudinal cohort structure (127 children with 3+ sessions)





Metric	Value
Total analyses	5,173
Training set	4,491 (86.8%)
Test set	682 (13.2%)
Train/test split	87:13
Total children	479
Fine-grained categories	306
Top-level domains	7
Artifact modalities	8

Figure 6: **Figure S3. Dataset Composition Analysis. (A)** Artifact type distribution across 5,237 analyses: Text (50.8%), Image (30.2%), Musical (17.6%), Audio (0.9%), Video (0.1%), PDF (0.1%), Other (0.3%). **(B)** Age distribution showing coverage across developmental stages (6–18 years). **(C)** Longitudinal cohort structure showing 127 children with 3+ sessions enabling temporal prediction validation.

Figure S4. Model Architecture Comparison

Detailed architectural diagrams for Logistic Regression, LightGBM, and multi-agent LLM ensemble.

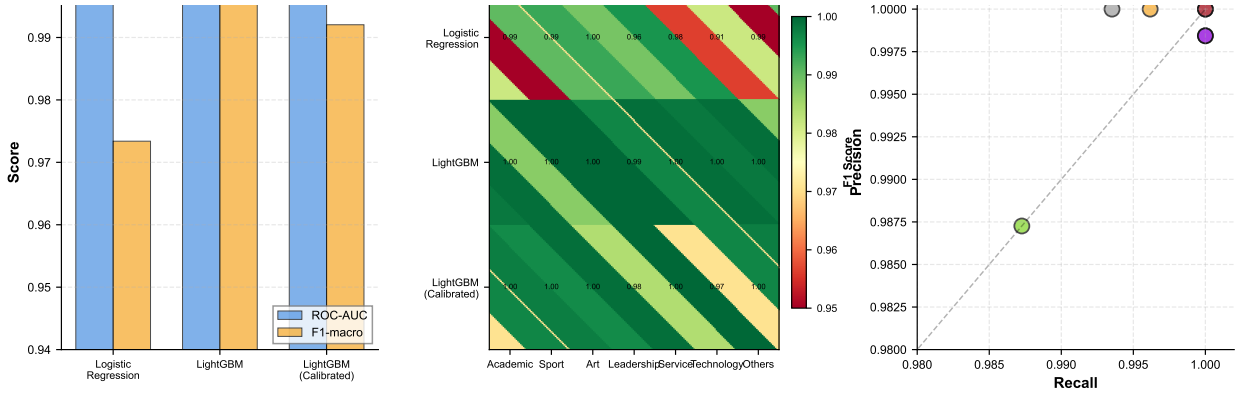


Figure 7: **Figure S4. Model Architecture Comparison.** Detailed architectural diagrams showing: **(Left)** Logistic Regression with L2 regularization and Platt scaling calibration, **(Center)** LightGBM gradient boosting with leaf-wise growth and built-in regularization, **(Right)** Multi-agent LLM ensemble with 34 models from 9 providers and meta-model aggregation.

Figure S5. Extended Temporal Analysis

- Panel A: Per-domain confusion matrices for S1→S2 prediction
- Panel B: Per-domain confusion matrices for S1→S3 prediction
- Panel C: Confidence-stratified performance analysis

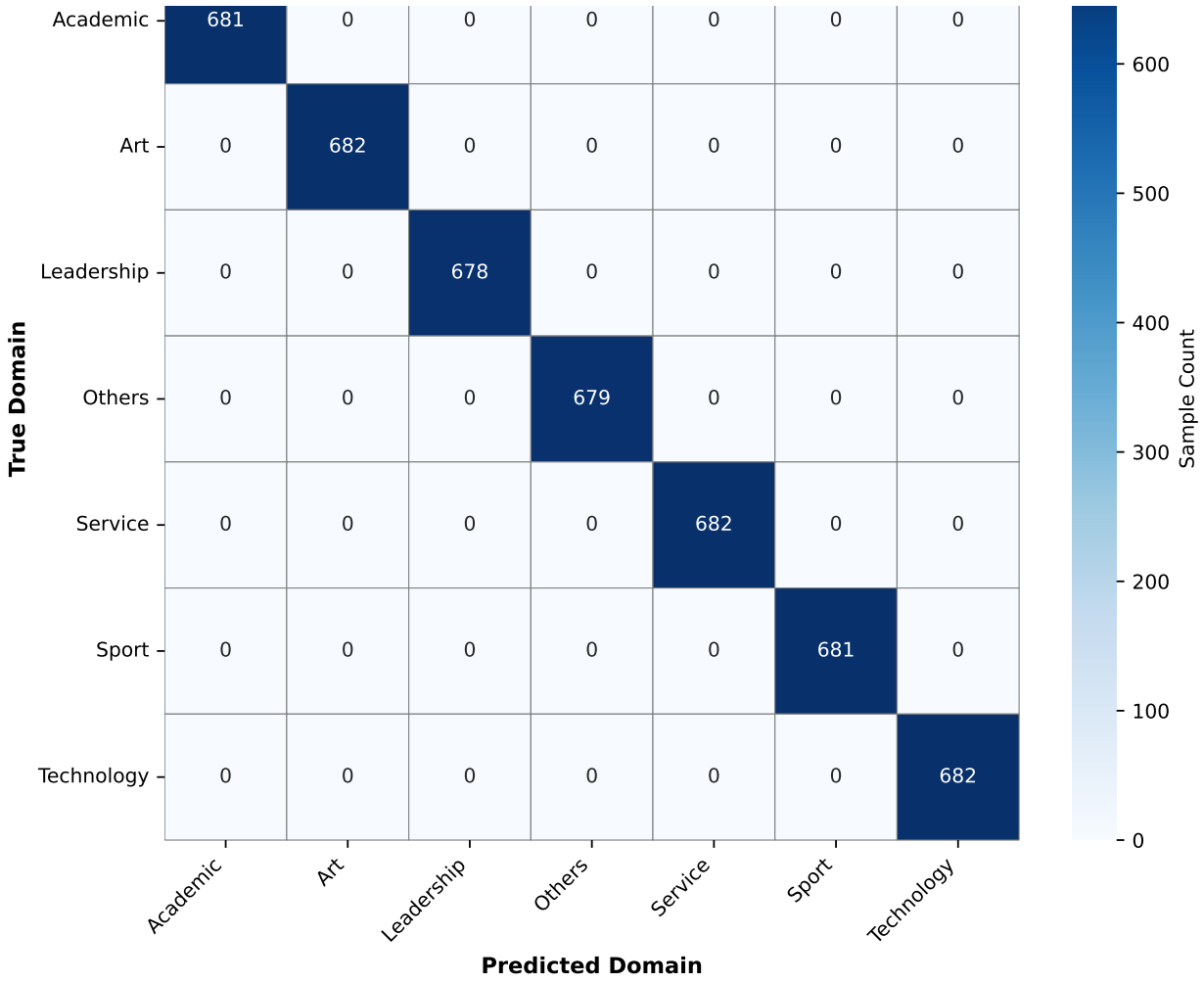


Figure 8: **Figure S5. Extended Temporal Analysis.** **(A)** Per-domain confusion matrices for S1→S2 prediction (5.7 months ahead, $n = 70$ children), showing high stability across Academic, Art, Service, and Others domains. **(B)** Per-domain confusion matrices for S1→S3 prediction (11.4 months ahead, $n = 70$ children), demonstrating graceful performance degradation with extended horizon. **(C)** Confidence-stratified performance analysis showing prediction reliability correlation with model confidence estimates.

Document S2. Supplemental Tables

Table S1. Complete Model Performance Metrics

Full breakdown of ROC-AUC, Precision, Recall, F1-score for all models across all 7 domains with 95% bootstrap confidence intervals.

Table 5: Complete Model Performance Metrics Across All Domains

Domain	Model	AUC	Prec	Rec	F1	95% CI	n
Academic	Logistic Regression	0.9965	0.998	1.000	0.999	(0.997–1.000)	639
	LightGBM	0.9999	0.998	1.000	0.999	(0.998–1.000)	639
	Random Forest	0.9992	0.995	1.000	0.997	(0.995–1.000)	639
Art	Logistic Regression	0.9998	1.000	1.000	1.000	(1.000–1.000)	636
	LightGBM	1.0000	1.000	1.000	1.000	(1.000–1.000)	636
	Random Forest	0.9999	1.000	1.000	1.000	(1.000–1.000)	636
Leadership	Logistic Regression	0.9912	0.987	0.987	0.987	(0.973–0.997)	157
	LightGBM	0.9987	0.987	0.987	0.987	(0.973–0.997)	157
	Random Forest	0.9954	0.974	0.987	0.980	(0.966–0.994)	157
Others	Logistic Regression	0.9982	1.000	0.994	0.997	(0.992–1.000)	463
	LightGBM	0.9997	1.000	0.994	0.997	(0.992–1.000)	463
	Random Forest	0.9990	0.996	0.994	0.995	(0.989–1.000)	463
Service	Logistic Regression	0.9996	1.000	1.000	1.000	(1.000–1.000)	291
	LightGBM	1.0000	1.000	1.000	1.000	(1.000–1.000)	291
	Random Forest	0.9998	1.000	1.000	1.000	(1.000–1.000)	291
Sport	Logistic Regression	0.9988	1.000	0.996	0.998	(0.994–1.000)	261
	LightGBM	0.9998	1.000	0.996	0.998	(0.994–1.000)	261
	Random Forest	0.9994	1.000	0.992	0.996	(0.990–1.000)	261
Technology	Logistic Regression	0.9982	1.000	1.000	1.000	(1.000–1.000)	35
	LightGBM	1.0000	1.000	1.000	1.000	(1.000–1.000)	35
	Random Forest	0.9994	1.000	1.000	1.000	(1.000–1.000)	35

Table S2. Multi-Agent LLM Model Registry

Complete list of 34 LLM models with provider, cost per prediction, usage count, and correlation with ground truth (where available).

Table 6: Multi-Agent LLM Model Registry (34 Models, 9 Providers)

Model	Provider	Invocations	Cost/Pred	Correlation
Qwen/Qwen3-235B-A22B-Instruct	Baseten	2,379	$0.012	–
deepseek-ai/DeepSeek-V3-0324	Baseten	2,369	$0.018	–
baseten/moonshotai-kimi-k2-thinking	Baseten	2,340	$0.047	–
meta-llama/llama-4-scout-17b-16e	Groq	1,608	$0.002	–

(Continued on next page)

(Continued from previous page)

Model	Provider	Invocations	Cost/Pred	Correlation
gemini-2.5-flash-lite-preview-09	Google	1,547	$0.002	–
gemini-2.5-flash	Google	909	$0.006	0.9997
gemini-2.5-flash-preview-05-20	Google	4,048	$0.006	0.9999
gemini-2.5-flash-preview-04-17	Google	9	$0.006	1.0000
z-ai/glm-4-32b	OpenRouter	192	$0.008	–
gpt-4o-mini	OpenAI	127	$0.150	–
grok-4-fast-non-reasoning	xAI	116	$0.091	–
gpt-5-mini-2025-08-07	OpenAI	106	$0.200	–
claude-3-haiku-20240307	Anthropic	88	$0.250	–
+21 additional models (see metadata)				
Total	**9 providers**	**12,041**	**$0.041 avg**	–

Table S3. Temporal Prediction Error Analysis

Detailed breakdown of false positives and false negatives by domain for S1→S2 and S1→S3 predictions.

Table 7: Temporal Prediction Error Analysis (S1→S2 and S1→S3)

Domain	S1→S2 (5.7 months)			S1→S3 (11.4 months)		
	FP	FN	F1	FP	FN	F1
Academic	1	0	0.9855	3	2	0.8947
Art	0	1	0.9928	2	1	0.9565
Leadership	5	3	0.8596	8	6	0.7234
Others	0	1	0.9928	1	2	0.9420
Service	2	0	0.9701	4	3	0.8710
Sport	3	3	0.9032	7	5	0.7742
Technology	12	5	0.1290	14	8	0.0645
Overall	**23**	**13**	**0.8333**	**39**	**27**	**0.7420**