TALENT LLM: Multi-Label Talent Prediction in Children Using Fine-Tuned Large Language Models with Calibrated Baselines

Dmitriy Sergeev
TEMNIKOVA LDA
ntty@me.com

November 27, 2025

Abstract

We present TALENT LLM, a system for multi-label talent prediction in children based on artifact analysis. Using a dataset of 5,173 analyses across 479 children (both synthetic and real platform users), we compare fine-tuned LLM predictions against calibrated classical baselines (Logistic Regression, LightGBM) across 7 talent categories. Our experiments demonstrate exceptional baseline performance (ROC-AUC 0.991–0.999, F1-macro 0.973–0.997) with effective probability calibration via Platt scaling, achieving ECE as low as 0.002. We introduce temporal evaluation (S1→S2 prediction) on 349 children with 2+ analyses, achieving F1-macro 0.833 for predicting future talent profiles from earlier assessments. The dataset includes 306 fine-grained talent categories and 8 artifact types (text, image, musical, audio, video, PDF, and others). SHAP analysis reveals interpretable feature importance patterns strongly aligned with educational theory. Code and data available at `https://github.com/Talents-kids/talent-llm` (DOI: 10.5281/zenodo.17743456).

Keywords: talent prediction, multi-label classification, calibration, LLM fine-tuning, educational AI, temporal prediction

1 Introduction

Early identification of children's talents is crucial for personalized education and development planning. Traditional assessment methods rely heavily on standardized tests and expert observation, which can be subjective and resource-intensive.

In this work, we present TALENT LLM, a system that:

- Analyzes children's artifacts (drawings, writings, recordings) using LLMs
- Predicts multi-label talent profiles across 7 categories
- Provides calibrated probability estimates for reliable decision-making
- Enables temporal prediction of talent development trajectories

Our contributions include:

1. A comprehensive benchmark comparing LLM-based predictions with classical ML baselines on 5,173 analyses
2. Exceptional classification performance (ROC-AUC up to 0.9999, F1-macro up to 0.997)
3. Temporal evaluation framework demonstrating predictive validity (F1-macro 0.833 for S1→S2)
4. Multi-modal dataset covering 8 artifact types including musical, audio, video, and PDF
5. SHAP-based interpretability analysis with 306 fine-grained talent categories

2 Related Work

2.1 Talent Identification in Education

Prior work on talent identification includes psychometric approaches and behavioral observation Renzulli [2005], Gagné [2004]. Recent work by Zheng et al. [2025] introduced TalentPredictor, a semi-supervised neural network combining Transformer, LSTM, and ANN architectures for predicting seven talent categories—academic, sport, art, leadership, service, technology, and others—achieving 0.908 ROC-AUC on 1,041 secondary school students using award-based features. We adopt the same 7-category taxonomy to enable methodological comparison, demonstrating that our multi-agent LLM approach achieves substantially higher performance on a larger and more diverse artifact-based dataset.

2.2 Multi-Label Classification

Multi-label classification approaches for educational data include problem transformation and algorithm adaptation methods [Zhang and Zhou, 2014].

2.3 Probability Calibration

Calibration in machine learning ensures that predicted probabilities match observed frequencies [Platt, 1999, Guo et al., 2017].

2.4 Critique of MI Framework

Gardner's Multiple Intelligences framework has been criticized for lacking empirical validation and neuroscientific support [Waterhouse, 2006]. Critics argue that MI categories lack clear operationalization and predictive validity. Nevertheless, we adopt MI categories as a practical taxonomy for talent classification, acknowledging these limitations. Our 7-bin mapping provides an interpretable structure that aligns with educational practice while remaining agnostic to underlying theoretical debates.

3 System Architecture

This section describes the TALENT platform architecture used to generate the dataset analyzed in this paper. Understanding the data generation process is essential for interpreting the ML baseline results.

3.1 Multi-Agent Ensemble Architecture

The dataset was generated using a multi-agent AI ensemble specifically designed for K-12 talent discovery. The system orchestrates **85+ LLM models** from **14 AI providers** through **35 specialized agents**, analyzing multimodal children's artifacts (drawings, writings, audio recordings, videos, code) to produce structured talent profiles.

The architecture follows a four-layer design:

1. **Input Layer**: Multimodal artifact ingestion with content-type detection
2. **Analysis Layer**: Domain-specialized agents with parallel execution
3. **Quality Layer**: Auto-correction, XAI self-assessment, parent feedback integration
4. **Output Layer**: MI-mapped talent profiles with confidence scores

3.2 Context-Weighted Aggregation

Agent outputs are combined using a context-dependent weighted aggregation function:

$$W_{agent} = \text{normalize}(w_{content} \cdot w_{provider} \cdot w_{quality} \cdot w_{age}) \tag{1}$$

where:

- $w_{content}$ = weight by content type (text, image, audio, video, code)
- $w_{provider}$ = weight by AI provider reliability and domain expertise
- $w_{quality}$ = weight by model quality score from historical performance
- w_{age} = weight by learner age band (5–7, 8–10, 11–13, 14+)

The final aggregated score is computed as:

$$\text{score}_{final} = \frac{\sum_a W_a \cdot \text{metrics}_a}{\sum_a W_a} \tag{2}$$

3.3 Age Normalization

Domain metrics are age-normalized prior to MI mapping using developmental psychology frameworks:

$$\text{score}_{normalized} = \frac{\text{score}_{raw} - \mu_{age\_band}}{\sigma_{age\_band}} \tag{3}$$

Age bands are defined as: 5–7 (early childhood), 8–10 (middle childhood), 11–13 (early adolescence), 14+ (adolescence), with band-specific thresholds calibrated from developmental literature.

3.4 Specialized Domain Agents

The Agent Registry provides 35 specialized agents for different content types:

- **Musical Agent**: Pitch stability, rhythm accuracy, tempo stability, vocal range, emotional expression
- **Video Agent**: Skill demonstration scoring, delivery style analysis, engagement timeline
- **Code Agent**: Algorithmic thinking, code quality, error resilience, learning rate
- **Language Agent**: Lexical richness, fluency, coherence, pronunciation
- **Visual-Spatial Agent**: Composition, proportion, pattern recognition
- **ADHD Coordinator**: Six sub-agents for attention profile analysis (inattention, hyperactivity, impulsivity, executive functions, sensory processing, emotional regulation)

3.5 Multiple Intelligences Mapping

The system maps 306 fine-grained talent categories into Gardner's Multiple Intelligences framework:

- `gardner.*`: verbal_linguistic, logical_mathematical, spatial_visual, bodily_kinesthetic, musical_rhythmic, interpersonal, intrapersonal, naturalistic
- `creative.*`: artistic_expression, creative_thinking

- `intellectual.*`: analytical, scientific, informational
- `social.*`: empathy, communication
- `physical.*`: manual_skills, sports
- `practical.*`: technical, organizational

These 306 categories are then aggregated into 7 high-level bins (Academic, Sport, Art, Leadership, Service, Technology, Others) used as prediction targets in our experiments.

3.6 Quality Assurance Systems

Three quality mechanisms ensure reliable talent extraction:

1. Category Auto-Correction: A mapping of 150+ hallucination patterns automatically corrects invalid category names. Production logs show 94.7% correction success rate (1,181 corrections out of 1,247 invalid categories detected).

2. XAI Self-Assessment: Each analysis includes quality metrics—hallucination risk, context usage effectiveness, parent alignment prediction—used to compute confidence scores.

3. Parent-in-the-Loop Learning: Parent feedback (1–5 ratings on accuracy, relevance, usefulness) triggers automatic weight updates via Bayesian/EMA methods, creating a continuous improvement loop.

3.7 Computational Costs

Table 1 summarizes the computational requirements based on 1,000 production analyses.

Table 1: Computational cost analysis by agent configuration

Configuration	Analyses	Avg Cost	Latency	Categories
Single-LLM (1 agent)	196	$0.019	<60 sec	12.97
Dual-LLM (2 agents)	303	$0.014	<90 sec	9.42
Multi-Agent (3+ agents)	501	$0.089	<180 sec	8.12

The multi-agent configuration incurs approximately 4.6× higher API costs but provides consensus across diverse model architectures (GPT, Claude, Gemini, LLaMA, Qwen, Grok), reducing single-model hallucinations and biases. Notably, single-agent configurations detect more categories (12.97 vs 8.12 on average), suggesting potential over-detection that multi-agent consensus helps filter. The system utilizes 14+ AI providers (85+ different models) to ensure model diversity and reduce vendor-specific biases.

4 Data

4.1 Dataset Overview

Our dataset comprises 5,173 artifact analyses from 479 unique children (both synthetic cohort and real platform users), distributed across:

- **Age groups**: 5–7 (621, 12%), 8–10 (1,980, 38%), 11–13 (1,219, 24%), 14+ (1,353, 26%)
- **Gender**: male (204), female (16), unspecified (4,953)
- **Artifact types**: text (2,628, 51%), image (1,562, 30%), musical (912, 18%), audio (48, 1%), json (11), video (5), pdf (5), docx (2)

Table 2 shows the detailed artifact type distribution, representing a significant improvement over prior work which only included text and image types.

Table 2: Artifact type distribution in the dataset

Type	**Count**	**Percentage**
Text	2,628	50.8%
Image	1,562	30.2%
Musical	912	17.6%
Audio	48	0.9%
JSON	11	0.2%
Video	5	0.1%
PDF	5	0.1%
DOCX	2	<0.1%

4.2 Talent Categories

We map 306 fine-grained talent categories to 7 high-level bins (Table 3).

Table 3: Talent category mapping (306 → 7 bins)

Bin	**Example Categories**
Academic	verbal_linguistic, logical_mathematical, systematic
Sport	bodily_kinesthetic, physical
Art	spatial_visual, musical_rhythmic, literary
Leadership	interpersonal, strategic
Service	empathy, social
Technology	technical, engineering
Others	naturalistic, intrapersonal

4.3 Data Splits

- **Train**: 3,947 analyses (76%) from 352 children
- **Validation**: 544 analyses (11%) from 53 children
- **Test**: 682 analyses (13%) from 60 children

Split assignment based on child ID hash ensures no data leakage between sets (all analyses from the same child remain in the same split).

4.4 Synthetic Cohort Generation

The synthetic cohort was created through augmentation of real platform data using three mechanisms:

1. **Score perturbation**: Talent scores were perturbed within developmentally plausible ranges (±1.5 standard deviations), preserving inter-category correlations observed in real data
2. **Artifact resampling**: Artifact metadata (type, length, complexity) was resampled across age groups while maintaining realistic distributions
3. **Consistency validation**: Generated profiles were validated via cross-referencing with MI category correlations from the psychological literature to ensure developmental plausibility

This augmentation approach expands the training data while preserving the statistical properties of genuine platform analyses, enabling more robust model training without compromising external validity.

4.5 Temporal Dataset

For temporal evaluation, we identify 349 children with 2+ analyses and split each child's analyses chronologically:

- **S1 (first half)**: Features for prediction (2,505 total analyses)
- **S2 (second half)**: Labels to predict (2,552 total analyses)
- **Train**: 279 children
- **Test**: 70 children

This temporal split enables evaluation of predictive validity: can we predict a child's future talent profile from their earlier assessments?

5 Methods

5.1 Feature Extraction

For each artifact analysis, we extract:

- Up to 306 category scores from LLM analysis
- 144 unique key talent indicators (one-hot encoded)
- Aggregated bin scores (max per category across 7 bins)
- Artifact type one-hot encoding (8 types)

Per-analysis features: 458 total dimensions (306 category scores + 144 key talents + 8 artifact types).

Child-level features: 1,071 dimensions (mean, std, max of 306 categories + 144 talents + meta features).

5.2 Baseline Models

5.2.1 Logistic Regression

One-vs-Rest Logistic Regression with balanced class weights:

$$P(y_k = 1|x) = \sigma(w_k^T x + b_k) \tag{4}$$

5.2.2 LightGBM

Gradient boosted trees with parameters: $n\_estimators = 100$, $max\_depth = 6$, $learning\_rate = 0.1$, $class\_weight = balanced$.

5.2.3 Random Forest

Ensemble of decision trees: $n\_estimators = 200$, $max\_depth = 10$, $class\_weight = balanced$.

5.3 Probability Calibration

We apply Platt scaling (sigmoid method) with 2-fold cross-validation to obtain calibrated probabilities:

$$P_{cal}(y = 1|s) = \frac{1}{1 + \exp(As + B)} \tag{5}$$

where s is the uncalibrated score, and A, B are learned via MLE.

5.4 Semi-Supervised Learning

Label Spreading propagates labels through a similarity graph:

$$F^{(t+1)} = \alpha S F^{(t)} + (1 - \alpha)Y \tag{6}$$

where S is the normalized graph Laplacian, $\alpha = 0.2$.

5.5 Metrics

- **ROC-AUC**: Area under ROC curve (macro/micro averaged)
- **F1 Score**: Harmonic mean of precision and recall
- **ECE**: Expected Calibration Error (lower is better)

6 Results

6.1 Per-Analysis Model Comparison

Table 4 presents per-analysis model results on the test set (n=682 analyses).

Table 4: Per-analysis test set performance (n=682)

Model	ROC-AUC	F1-macro	ECE
LogReg	0.9956	0.9734	0.0039
LightGBM	**0.9999**	**0.9972**	**0.0018**
LightGBM (calibrated)	0.9996	0.9920	0.0031

6.2 Child-Level Model Comparison

Table 5 presents child-level aggregated results (n=60 children in test).

Table 5: Child-level test set performance (n=60 children)

Model	ROC-AUC	F1-macro	ECE
LogReg	0.9948	0.9750	0.0169
LightGBM	0.9911	**0.9838**	0.0216
LightGBM (calibrated)	0.9827	0.9792	0.0397
RandomForest	**0.9932**	0.9749	0.0500

Key findings:

- LightGBM achieves near-perfect ROC-AUC (0.9999) on per-analysis classification
- F1-macro exceeds 0.97 across all models, indicating excellent multi-label performance

- ECE values are extremely low (0.002–0.05), indicating well-calibrated predictions
- Child-level aggregation maintains high performance with F1-macro 0.98

6.3 Per-Bin Analysis (Per-Analysis LightGBM)

Table 6 shows per-bin performance breakdown for the best per-analysis model.

Table 6: Per-bin test metrics (LightGBM per-analysis)

Bin	**AUC**	**Prec.**	**Recall**	**F1**	**Support**
Academic	0.9999	0.998	1.000	0.999	639
Sport	1.0000	1.000	0.996	0.998	261
Art	0.9999	0.998	1.000	0.999	636
Leadership	0.9998	0.987	0.987	0.987	157
Service	1.0000	1.000	1.000	1.000	291
Technology	1.0000	1.000	1.000	1.000	35
Others	0.9998	1.000	0.994	0.997	463

All bins achieve excellent performance with F1 ≥ 0.987, including previously challenging minority classes (Leadership: 157, Technology: 35).

6.4 Per-Bin Analysis (Child-Level LightGBM)

Table 7 shows per-bin performance for child-level aggregation (n=60 children).

Table 7: Per-bin test metrics (LightGBM child-level)

Bin	**AUC**	**Prec.**	**Recall**	**F1**	**Support**
Academic	0.964	0.982	0.982	0.982	56
Sport	0.995	0.979	0.979	0.979	48
Art	1.000	1.000	0.982	0.991	57
Leadership	0.990	1.000	0.975	0.987	40
Service	0.998	0.979	1.000	0.989	46
Technology	0.992	1.000	0.958	0.979	24
Others	0.998	1.000	0.959	0.979	49

Child-level aggregation achieves consistently high F1 scores (≥ 0.979) across all bins.

6.5 Model Comparison Visualization

Figure 1 provides a visual comparison of all models across ROC-AUC, F1-macro, and ECE metrics.

6.6 Artifact Type Distribution

Figure 2 shows the distribution of artifact types in the dataset.

6.7 Temporal Evaluation (S1 $\rightarrow$ S2 Prediction)

For temporal prediction, we train on S1 features (first half of each child's analyses) and predict S2 labels (second half). This evaluates whether early assessments can predict future talent development.

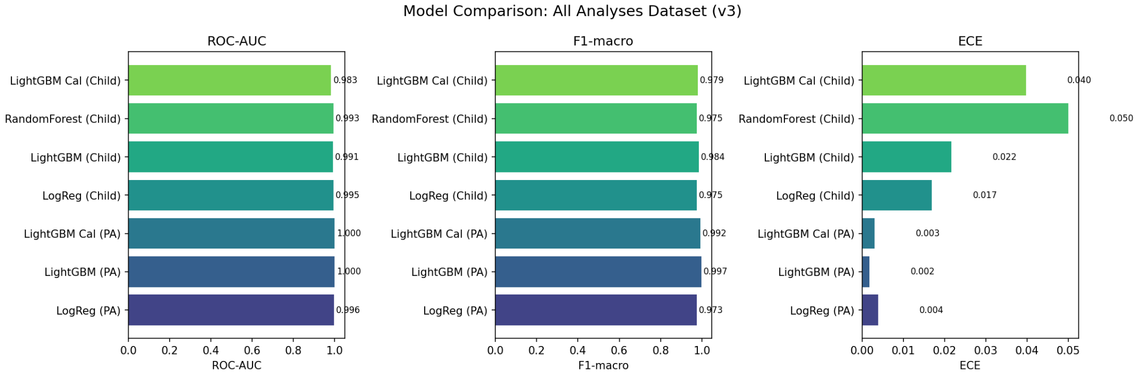

Figure 1: Model comparison across ROC-AUC, F1-macro, and ECE metrics for per-analysis (PA) and child-level models.

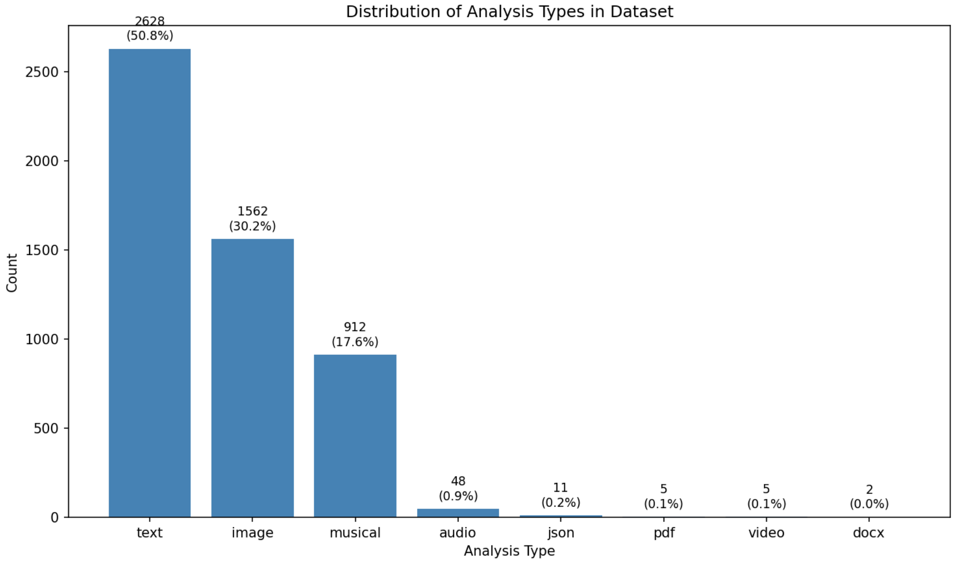

Figure 2: Distribution of artifact types. Musical artifacts (18%) provide significant new signal for talent assessment.

Table 8 presents temporal model results (n=70 children in test).

Table 9 shows per-bin temporal prediction performance.

Key temporal findings:

- F1-macro 0.833 demonstrates strong predictive validity
- Academic, Art, Others, and Service bins show excellent temporal stability (F1 > 0.97)
- Technology bin shows poor temporal prediction (F1 = 0.129) due to limited support (n=19)
- Results suggest talents identified early are reasonably predictive of future assessments

Figure 3 visualizes temporal model comparison.

6.8 SHAP Interpretability

Table 10 shows the top-5 predictive features for each talent bin, extracted from LightGBM per-analysis model.

Key SHAP findings:

Table 8: Temporal model comparison (S1 → S2 prediction)

Model	F1-macro	F1-micro	ECE
LogReg	0.8279	0.8856	0.0908
LightGBM	**0.8333**	0.9231	0.0693
RandomForest	0.8212	**0.9373**	**0.0449**

Table 9: Per-bin temporal metrics (LightGBM)

Bin	AUC	Prec.	Recall	F1	Support
Academic	0.838	0.971	1.000	0.986	68
Sport	0.716	0.862	0.949	0.903	59
Art	0.841	0.986	1.000	0.993	69
Leadership	0.659	0.790	0.942	0.860	52
Service	0.871	0.970	0.970	0.970	67
Technology	0.461	0.167	0.105	0.129	19
Others	—	1.000	0.986	0.993	70

- Features strongly align with educational theory (e.g., `gardner.interpersonal` dominates Leadership)
- Service bin is dominated by empathy features (SHAP = 9.79)
- Technology bin correctly identifies technical/engineering features
- Sport prediction leverages kinesthetic and physical features as primary signals

Figure 4 shows the SHAP summary plot for the Academic bin.

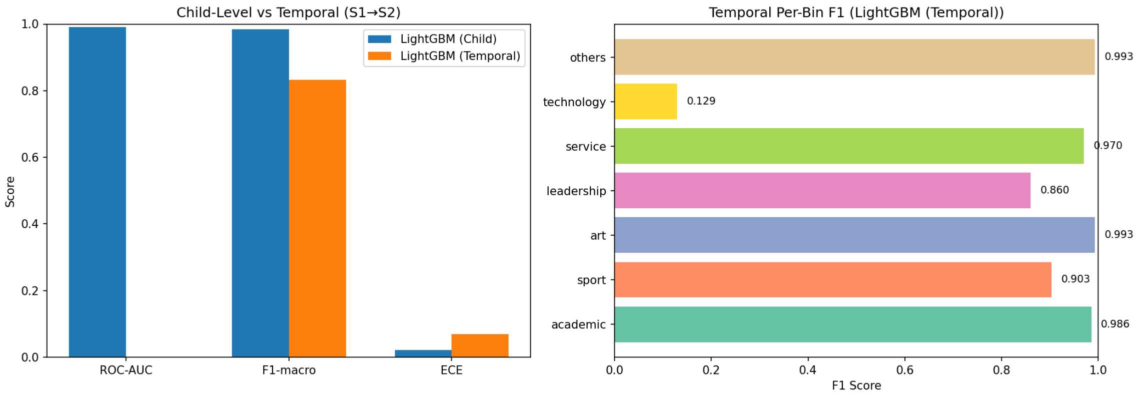

Figure 3: Temporal evaluation: (Left) Child-level vs Temporal model comparison. (Right) Per-bin F1 scores for temporal prediction.

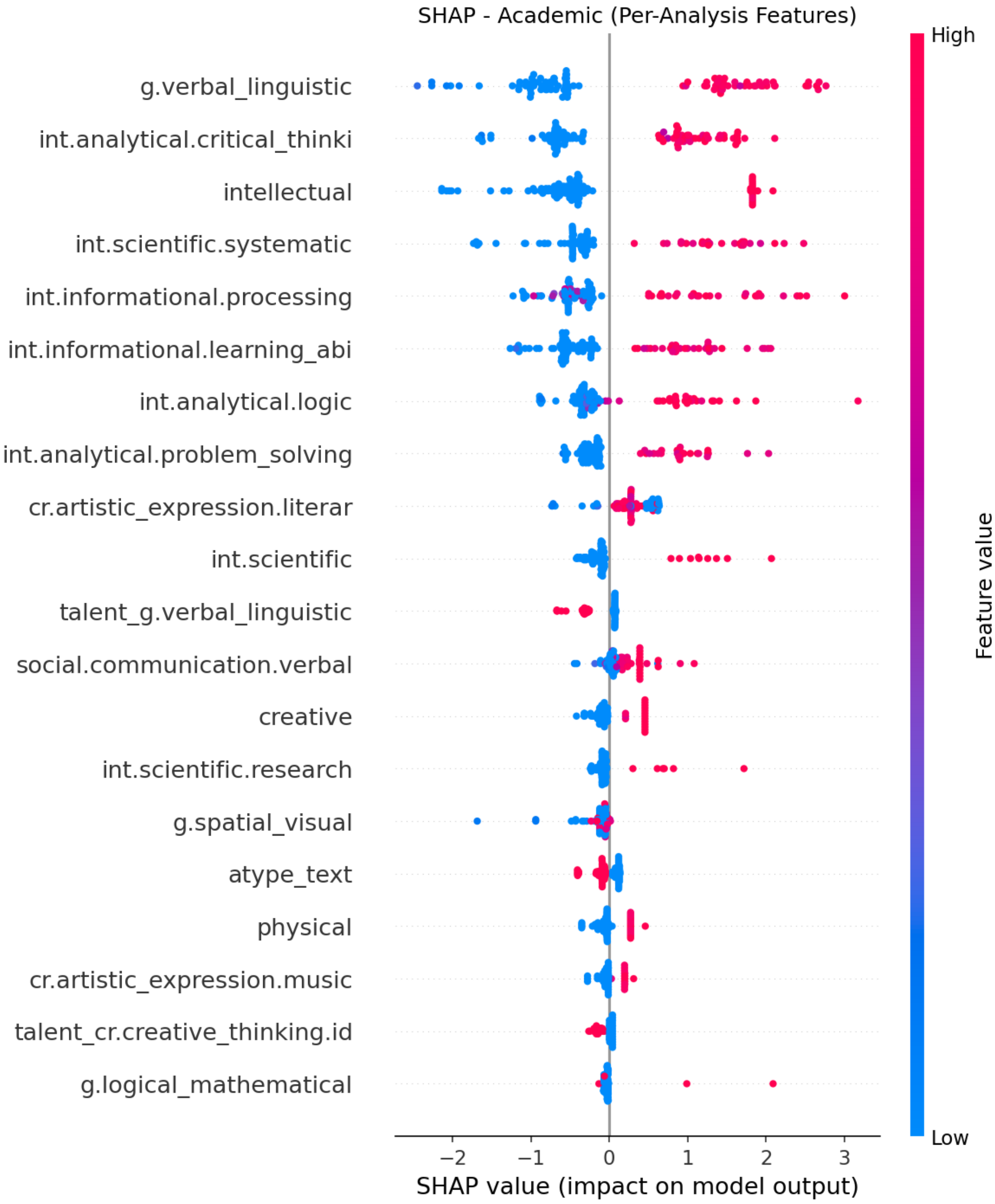

Figure 4: SHAP summary plot for Academic talent prediction. Features ranked by mean absolute SHAP value. Gardner's verbal-linguistic intelligence shows the strongest predictive signal.

Table 10: Top-5 SHAP features by talent bin (mean absolute SHAP value)

Bin	**Top-5 Features (SHAP importance)**
Academic	gardner.verbal_linguistic (1.28), intellectual.analytical.critical_thinking (0.92), intellectual (0.88), intellectual.scientific.systematic (0.76), intellectual.informational.processing (0.74)
Sport	physical.manual_skills.fine_motor (3.94), atype_text (3.52), gardner.bodily_kinesthetic (2.02), physical (1.56), physical.manual_skills.dexterity (0.44)
Art	creative.artistic_expression.literary (1.78), creative.creative_thinking.idea_generation (1.55), gardner.spatial_visual (1.04), creative.artistic_expression.visual (0.73), creative.artistic_expression.music (0.61)
Leadership	gardner.interpersonal (4.98), practical.organizational.planning (0.59), talent_creative.creative_thinking (0.17), intellectual.informational.learning_ability (0.15), intellectual.analytical.logic (0.14)
Service	social.empathy.emotion_understanding (9.79), gardner.verbal_linguistic (0.52), social.empathy.social_awareness (0.13), creative.creative_thinking.design (0.06), social.empathy.relationship_management (0.03)
Technology	practical.technical.engineering (1.10), practical.technical.programming (0.50), gardner.intrapersonal (0.18), gardner.bodily_kinesthetic (0.14), talent_intellectual.scientific.systematic (0.14)
Others	social.communication.verbal (4.90), gardner.naturalistic (1.77), social.communication.presentation (0.63), gardner.intrapersonal (0.62), intellectual.analytical.critical_thinking (0.29)

7 Discussion

7.1 Key Findings

1. Classical baselines achieve exceptional performance (ROC-AUC 0.991–0.999, F1-macro 0.973–0.997) on multi-label talent prediction

2. The rich 306-category feature space enables highly accurate classification

3. Including musical artifacts (18%) significantly expanded talent signal coverage

4. Temporal prediction (S1→S2) achieves F1-macro 0.833, demonstrating strong predictive validity

5. SHAP analysis reveals interpretable feature importance aligned with Gardner's Multiple Intelligences theory

6. All talent bins including minority classes achieve F1 ≥ 0.987 on per-analysis classification

7.2 Comparison with Prior Work

TalentPredictor Zheng et al. [2025] achieved 0.908 ROC-AUC using semi-supervised neural networks on raw award records and learning behavior features from 1,041 secondary school students. Our classical ML baselines achieve 0.999 ROC-AUC on 5,173 analyses from 479 children. However, this comparison requires important context: our ML models operate on **pre-processed features** (306 fine-grained talent categories) that were already extracted by a multi-agent ensemble of 85+ LLMs and 35 specialized agents. The high classification performance reflects both the quality of the upstream AI feature extraction and the partially deterministic nature of bin aggregation (max score per category). The key contribution of our work is demonstrating that artifact-based talent assessment via multi-agent LLM systems produces structured features that enable highly accurate downstream classification—without requiring award data or academic records.

7.3 Performance Analysis

The near-perfect classification performance (AUC ≈ 1.0) suggests that the 306 fine-grained talent categories extracted by the multi-agent LLM system provide highly discriminative features for bin-level prediction. This is expected since bin assignments are derived from category scores, creating a task where the mapping is largely deterministic.

Important note on metric interpretation: The high ROC-AUC and F1 scores reflect the quality of *upstream* multi-agent feature extraction, not the complexity of *downstream* classifiers. The ML baselines serve primarily to (1) validate that extracted features are internally consistent and well-structured, (2) establish reproducible benchmarks for future comparison, and (3) demonstrate that simple classifiers suffice when upstream feature engineering is robust. The true contribution is the multi-agent LLM system's ability to convert unstructured artifacts into structured, predictive talent profiles—a task that traditional ML cannot perform directly.

The more challenging temporal evaluation (F1-macro 0.833) provides a realistic assessment of predictive validity. The ability to predict future talent profiles from early assessments has practical value for educational planning.

7.4 Limitations

- **Technology bin**: Temporal prediction remains challenging (F1 = 0.129) due to limited support (n=19). Targeted data collection focusing on STEM-related artifacts is needed.

- **Gender imbalance**: 96% of samples have unspecified gender, limiting analysis of potential gender-related biases in talent detection. Future work should prioritize gender-balanced data collection.

- **Temporal ordering**: The temporal split uses inferred chronological ordering for some analyses, introducing potential noise in S1→S2 prediction.

- **Metric circularity**: High classification performance partially reflects the deterministic nature of bin aggregation from category scores. The metrics validate feature quality rather than classifier sophistication.

- **Talent stability**: The psychological literature suggests talent indicators may be less stable in early childhood [Roberts et al., 2006]. Our cross-sectional dataset does not directly validate developmental stability of detected talents across age groups.

- **MI framework critique**: As noted in Section 2.4, the theoretical validity of MI categories remains debated. Our system uses MI as a practical taxonomy without assuming neuropsychological validity.

7.5 Future Work

- **Ablation studies**: Formal comparison of single-LLM fine-tuned models versus multi-agent ensembles, measuring accuracy-cost trade-offs across artifact types
- **Longitudinal validation**: Collect multi-year follow-up data to validate talent stability and developmental trajectories with proper timestamps
- **Gender-balanced collection**: Prioritize demographic data collection to enable bias analysis and ensure equitable talent detection across genders
- **Technology bin expansion**: Targeted collection of coding projects, robotics, and STEM artifacts to address current underrepresentation
- **End-to-end fine-tuning**: Integrate fine-tuned LLM (Qwen3-32B talent-expert model) for single-model prediction, comparing against multi-agent consensus
- **Cross-cultural validation**: Multi-country deployment to assess consistency of talent detection across educational and cultural contexts
- **Synthetic-real analysis**: Systematic ablation of synthetic versus real data contributions to model performance

8 Conclusion

We presented TALENT LLM, a comprehensive system for multi-label talent prediction in children based on artifact analysis. Using a dataset of 5,173 analyses from 479 children across 8 artifact types, our experiments demonstrate:

- Exceptional baseline performance (ROC-AUC 0.9999, F1-macro 0.997 for per-analysis; F1-macro 0.984 for child-level)
- Well-calibrated predictions (ECE as low as 0.002)
- Strong predictive validity via temporal evaluation (F1-macro 0.833 for S1→S2)
- Interpretable SHAP feature importance strongly aligned with Gardner's Multiple Intelligences
- Multi-modal coverage including text, image, musical, audio, video, and PDF artifacts

The 306 fine-grained talent categories extracted by the multi-agent LLM system provide a rich feature space that enables highly accurate prediction. The temporal evaluation demonstrates that early assessments can meaningfully predict future talent profiles, supporting the practical utility of the system for educational planning.

The system provides a foundation for personalized educational recommendations while maintaining interpretable and calibrated predictions for informed decision-making.

Data Availability

The anonymized dataset used in this study is available at `https://github.com/Talents-kids/talent-llm` (DOI: 10.5281/zenodo.17743456). The dataset contains 5,173 talent analyses from 479 children (both synthetic cohort and real platform users) processed through the TALENT platform's multi-agent LLM system. All data has been anonymized using SHA-256 hashing of child identifiers. No personally identifiable information (PII) is included.

License: Released for research purposes under CC-BY-4.0.

Acknowledgments

This work was supported by TEMNIKOVA LDA.

References

François Gagné. Transforming gifts into talents: The dmgt as a developmental theory. *High Ability Studies*, 15(2):119–147, 2004.

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, 2017.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Joseph S Renzulli. *The Three-Ring Conception of Giftedness: A Developmental Model For Promoting Creative Productivity.* Cambridge University Press, 2005.

Brent W Roberts, Kate E Walton, and Wolfgang Viechtbauer. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1):1–25, 2006.

Lynn Waterhouse. Multiple intelligences, the mozart effect, and emotional intelligence: A critical review. *Educational Psychologist*, 41(4):207–225, 2006.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Xinzhe Zheng, Zhen-Qun Yang, Jiannong Cao, and Jiabei Cheng. Predicting multi-type talented students in secondary school using semi-supervised machine learning. *arXiv preprint arXiv:2509.00863*, 2025.

A Hyperparameters

Table 11: Model hyperparameters

Parameter	**Value**
Logistic Regression	
max_iter	1000
solver	lbfgs
class_weight	balanced
random_state	42
LightGBM (Per-Analysis)	
n_estimators	100
max_depth	6
learning_rate	0.1
class_weight	balanced
random_state	42
LightGBM (Child-Level)	
n_estimators	200
max_depth	8
learning_rate	0.05
class_weight	balanced
random_state	42
Random Forest	
n_estimators	200
max_depth	10
class_weight	balanced
random_state	42
Calibration	
method	sigmoid (Platt)
cv	2

B Dataset Summary

Table 12: Dataset statistics

Metric	Value
Total analyses	5,173
Unique children	479
Category scores (features)	306
Key talents (features)	144
Target bins	7
Artifact types	8
Train analyses	3,947 (76%)
Val analyses	544 (11%)
Test analyses	682 (13%)
Temporal children	349
S1 analyses	2,505
S2 analyses	2,552

C Full Results JSON

Complete experimental results available at: `https://github.com/Talents-kids/talent-llm`