Teacher vs Student Model Comparison

Qwen 72B vs Qwen 1.5B on Silver Dataset

Performance Analysis & Knowledge Distillation Assessment

Executive Summary

This report presents a comprehensive comparison between the Qwen2.5-72B teacher model and Qwen2.5-1.5B student model on the silver dataset containing 100 samples. The analysis evaluates the effectiveness of knowledge distillation and assesses whether the 97.9% smaller student model can maintain acceptable performance for production deployment in UPSC content classification tasks.

Key Findings

Metric	Teacher (72B)	Student (1.5B)
Overall Accuracy	**80.00%**	**50.00%**
Precision	50.00%	32.69%
Recall	83.33%	53.12%
F1-Score	62.50%	40.48%
GS Paper Accuracy	80.00%	41.18%
Avg Time/Sample	4.67s	2.09s
Model Size	72B params	1.5B params

Critical Analysis

Performance Gap

The student model demonstrates a significant performance degradation across all metrics:

• Accuracy Drop: -30.00 percentage points (80% → 50%)
• Precision Drop: -17.31 percentage points (50% → 32.69%)
• Recall Drop: -30.21 percentage points (83.33% → 53.12%)
• F1-Score Drop: -22.02 percentage points (62.5% → 40.48%)• GS Paper Accuracy Drop: -38.82 percentage points (80% → 41.18%)

Confusion Matrix Comparison

Teacher Model (72B) - 30 Samples:

	Pred Neg	Pred Pos
Actual Neg	**19 (TN)**	5 (FP)
Actual Pos	1 (FN)	**5 (TP)**

Student Model (1.5B) - 100 Samples:

	Pred Neg	Pred Pos
Actual Neg	**33 (TN)**	**35 (FP)**
Actual Pos	**15 (FN)**	**17 (TP)**

GS Paper Classification Performance

GS Paper	Teacher Accuracy	Student Accuracy	Gap
GS2	75.0% (3/4)	41.2% (7/17)	-33.8%
GS3	50.0% (1/2)	0.0% (0/15)	**-50.0%**

Problem Analysis

Critical Issues Identified

1. Excessive False Positives
The student model flagged 35 irrelevant articles as relevant (compared to 5 for teacher). This represents a 7x increase in false positives, indicating the model has learned to over-classify content as relevant.

2. High False Negative Rate
With 15 false negatives (vs 1 for teacher), the student model missed 46.9% of relevant content. This is unacceptable for a content filtering system meant to support UPSC preparation.

3. Complete GS3 Failure

The student model achieved 0% accuracy on GS3 (Economy, Environment, Science & Technology) classification. This suggests the model fundamentally lacks understanding of this category's characteristics.

4. Confidence Calibration Issue
Despite poor performance, the student model reports high confidence (87.3% average). This indicates severe miscalibration where the model is confidently wrong.

Speed vs Accuracy Trade-off

While the student model offers significant computational advantages:
 • 2.2x faster inference (2.09s vs 4.67s per sample)
• 97.9% smaller model size (1.5B vs 72B parameters)
• Can run on single GPU (14.56 GB available, only 2.88 GB used) These benefits are completely overshadowed by the 30 percentage point accuracy drop. The model's 50% accuracy is barely better than random guessing, making it unsuitable for production deployment.

Root Cause Analysis

Why Knowledge Distillation Failed

1. No Actual Distillation Applied
The current 'student' model is simply the pre-trained Qwen 1.5B without any knowledge distillation training. It has not learned from the teacher model's outputs, logits, or decision patterns.

2. Insufficient Model Capacity
1.5B parameters may be too small to capture the nuanced understanding required for UPSC content classification. The 48x reduction in model size (72B → 1.5B) is likely too aggressive.

3. Domain-Specific Knowledge Gap
UPSC content classification requires understanding of Indian governance, policy frameworks, and examination patterns. The smaller model lacks this specialized knowledge without proper fine-tuning.

4. Training Data Mismatch
The pre-trained Qwen 1.5B was not trained specifically on Indian current affairs, newspaper content, or UPSC-style material, creating a significant distribution gap.

Recommendations

Immediate Actions Required

1. Implement Proper Knowledge Distillation

Train the student model using:

- Teacher model logits as soft labels
- Temperature scaling (T=2-4) to soften distributions
- Combined loss: α·KL-divergence + (1-α)·cross-entropy
- Use teacher's confidence scores for sample weighting

2. Consider Intermediate Model Size

Test Qwen 7B or 14B as an intermediate option:

- Still 5-10x smaller than 72B
- Better capacity for complex classification
- More likely to preserve teacher performance

3. Domain-Specific Fine-Tuning

Create a curated dataset of 5,000-10,000 labeled examples:

- Balanced across GS1, GS2, GS3, GS4
- Include hard negatives (non-UPSC content)
- Use teacher annotations as gold standard
- Augment with similar content variations

4. Two-Stage Distillation

Implement progressive distillation:

- Stage 1: Distill 72B → 7B (closer in capacity)
- Stage 2: Distill 7B → 1.5B (incremental compression)
- Validate at each stage before proceeding

Alternative Approaches

1. Ensemble Method

Use multiple smaller models:

- Train 3-5 specialized 1.5B models (one per GS paper + relevance)
- Combine predictions with voting or weighted average
- May achieve better accuracy than single larger model

2. Hybrid System

Combine student model with rule-based filters:

- Use student for initial fast filtering

- Apply keyword/pattern rules for GS paper classification
- Teacher model validates only uncertain cases

3. LoRA Fine-Tuning

Apply Low-Rank Adaptation:

- Fine-tune only small adapter layers (~0.1% of parameters)
- Much faster and cheaper than full fine-tuning
- Can achieve significant improvements with limited data

Success Criteria for Next Iteration

Minimum Acceptable Performance:

Metric	Target
Relevance Accuracy	**≥ 75%**
Precision	≥ 45%
Recall	≥ 75%
GS Paper Accuracy	≥ 70%
Maximum Accuracy Gap from Teacher	≤ 10%

Conclusion

Current Status: NOT READY FOR PRODUCTION

The student model in its current form cannot replace the teacher model for UPSC content classification. The 30 percentage point accuracy drop represents a catastrophic loss in performance that would severely compromise the quality of content curation for UPSC aspirants.

Critical Findings:

- 50% accuracy is barely better than random guessing
- Complete failure on GS3 classification (0% accuracy)
- 7x increase in false positives would flood users with irrelevant content
- High false negative rate means missing nearly half of relevant content• Speed improvements are meaningless without acceptable accuracy

Next Steps:

1. Implement proper knowledge distillation training
2. Consider intermediate model size (7B or 14B)
3. Create domain-specific fine-tuning dataset
4. Re-evaluate after applying recommended improvements5. Continue using 72B teacher model for production until student achieves ≥75% accuracy

The goal of knowledge distillation is to create an efficient model that maintains performance. This experiment clearly demonstrates that simply using a smaller pre-trained model without distillation training is insufficient. Proper implementation of the recommended strategies is essential before the student model can be considered production-ready.

Appendix: Technical Details

Test Configuration

Parameter	Value
Teacher Samples	30
Student Samples	100
Dataset	Silver Dataset
Teacher Model	Qwen2.5-72B-Instruct
Student Model	Qwen2.5-1.5B-Instruct
Student Inference Device	Tesla T4 GPU (14.56 GB)
Temperature	0.1
Evaluation Date	February 15, 2026

Report Generated: February 15, 2026

Files: qwen72b_teacher_results (30 samples) | qwen1.5b_student_results (100 samples)