NEURAL INFORMATION PROCESSING SYSTEMS

Closed-Task Validation: A More Robust and Efficient Proxy for Guiding VLM Training

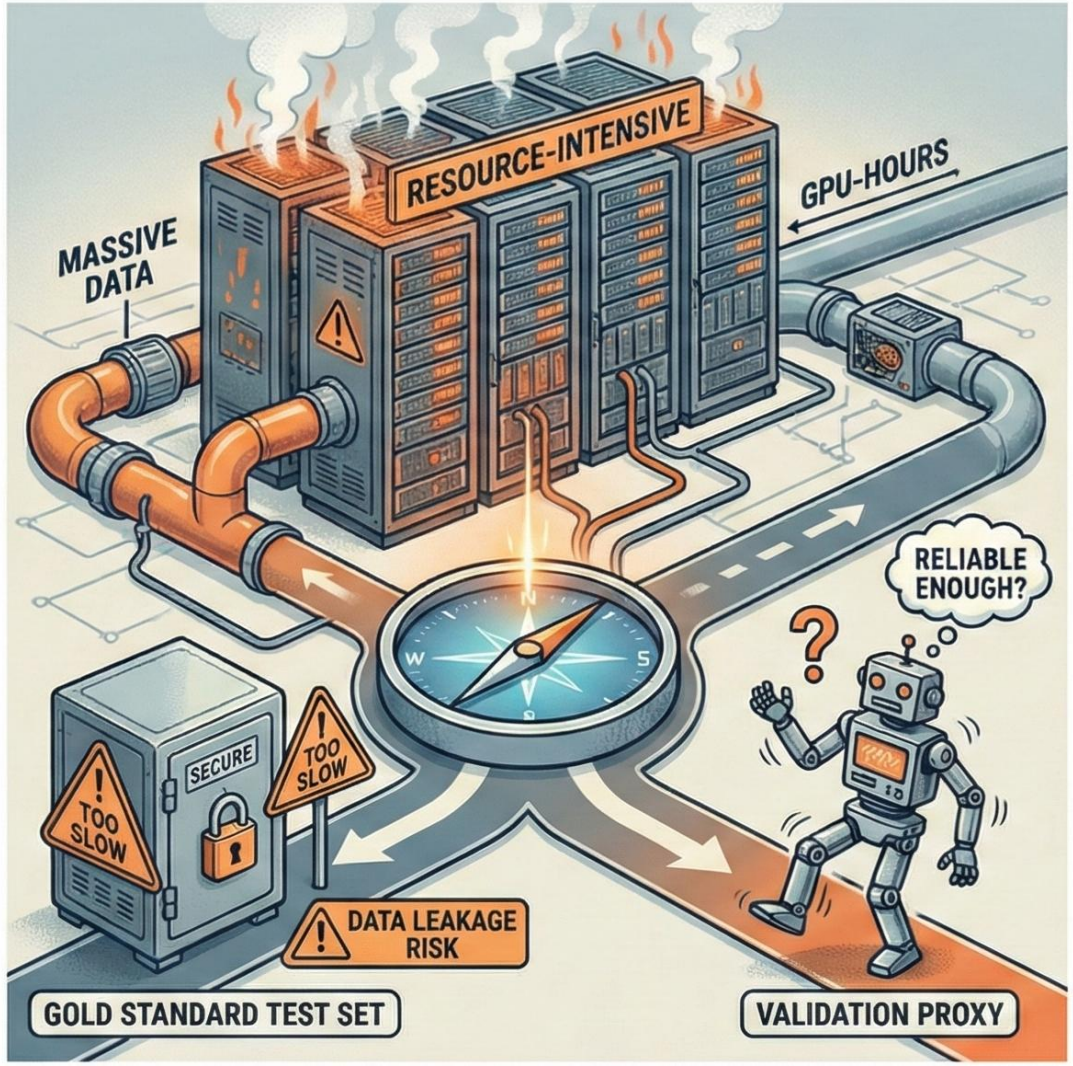
Enci Zhang[1,2], Zongqiang Zhang[1], Jiahao Xie[1], Ruiqi Lu[1], Boyan Zhou[1], Cheng Yang[1]*

[1]ByteDance
[2]Peking University

北京大学 PEKING UNIVERSITY

ByteDance

Background - The High Stakes of VLM Training



Resource-Intensive:

Training modern VLMs (e.g., Qwen-VL) consumes massive data (millions of samples) and GPU-hours.

The Role of Validation:

Acts as a "Compass" for hyperparameter tuning and checkpoint selection.

Critical: We cannot afford to evaluate on the full "Gold Standard" test set frequently (Too slow & Data leakage risk).

The Problem:

We rely on Validation Proxies.

Question: **Are our current proxies (Open-Ended Evaluation) reliable enough?**

Experimental Setup: Benchmarking the "Compass"

Dataset	Samples	Open-Task Metric	Closed-Task Metric
ChartQA[31]	450	Exact Match (Acc)	Exact Match (Acc)
InfoVQA[32]	450	ANLS	
DocVQA[33]	450	ANLS	
MathVista[34]	250	Exact Match (Acc)	
MMVet[35]	180	LLM-as-Judge	
VizWiz[36]	450	Exact Match (Acc)	

Table 1: Composition of the Paired Validation Set. Details the 2,230 samples drawn from six diverse benchmarks. Crucially, this set allows for a paired comparison, where Open-Task baselines use varied metrics while the Closed-Task paradigm uses a uniform Exact Match on probability.

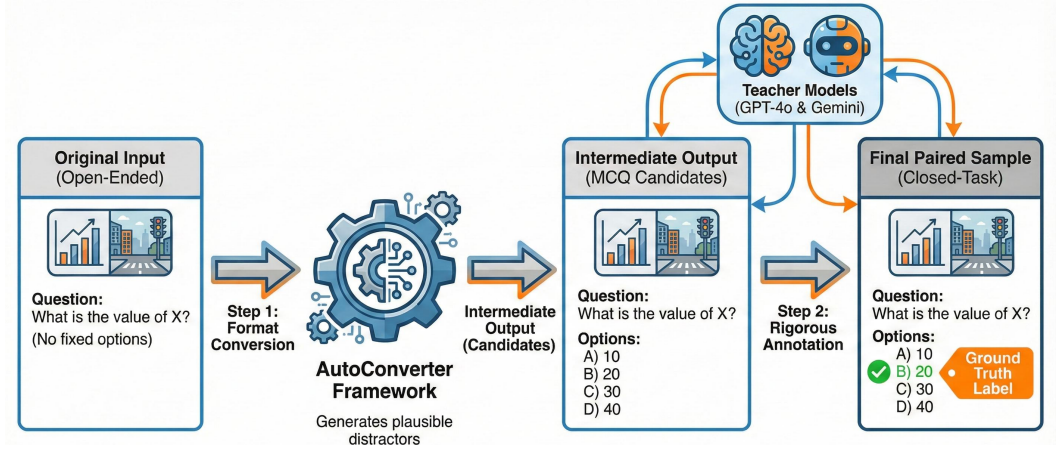


Figure 1: The Closed-Task Data Construction Pipeline. Step 1 involves converting the format via AutoConverter to generate plausible options. Step 2 ensures high quality through rigorous annotation by advanced LLMs to establish the ground truth.

Training Context

Compute: 16 × A100 GPUs, trained for 3,100 steps.
Data: 5 Million+ multimodal instruction samples.

Validation Set Construction

Dataset: 2,230 paired samples from 6 benchmarks
Construction: Converted via AutoConverter and annotated by GPT-4o & Gemini to ensure high quality.
Baseline: Open-Ended Generation (Text Output).
Ours: Closed-Task Validation (Multiple-Choice Probability).

Evaluation Framework & Gold Standard

The Comparison: Open-Task vs. Closed-Task (Token Prob).
Gold Standard: Full evaluation on a 39.4k held-out test set.
Goal: Establish the True Performance Curve to benchmark the reliability (correlation) of each proxy.

Finding 1: High Stability & Strong Correlation

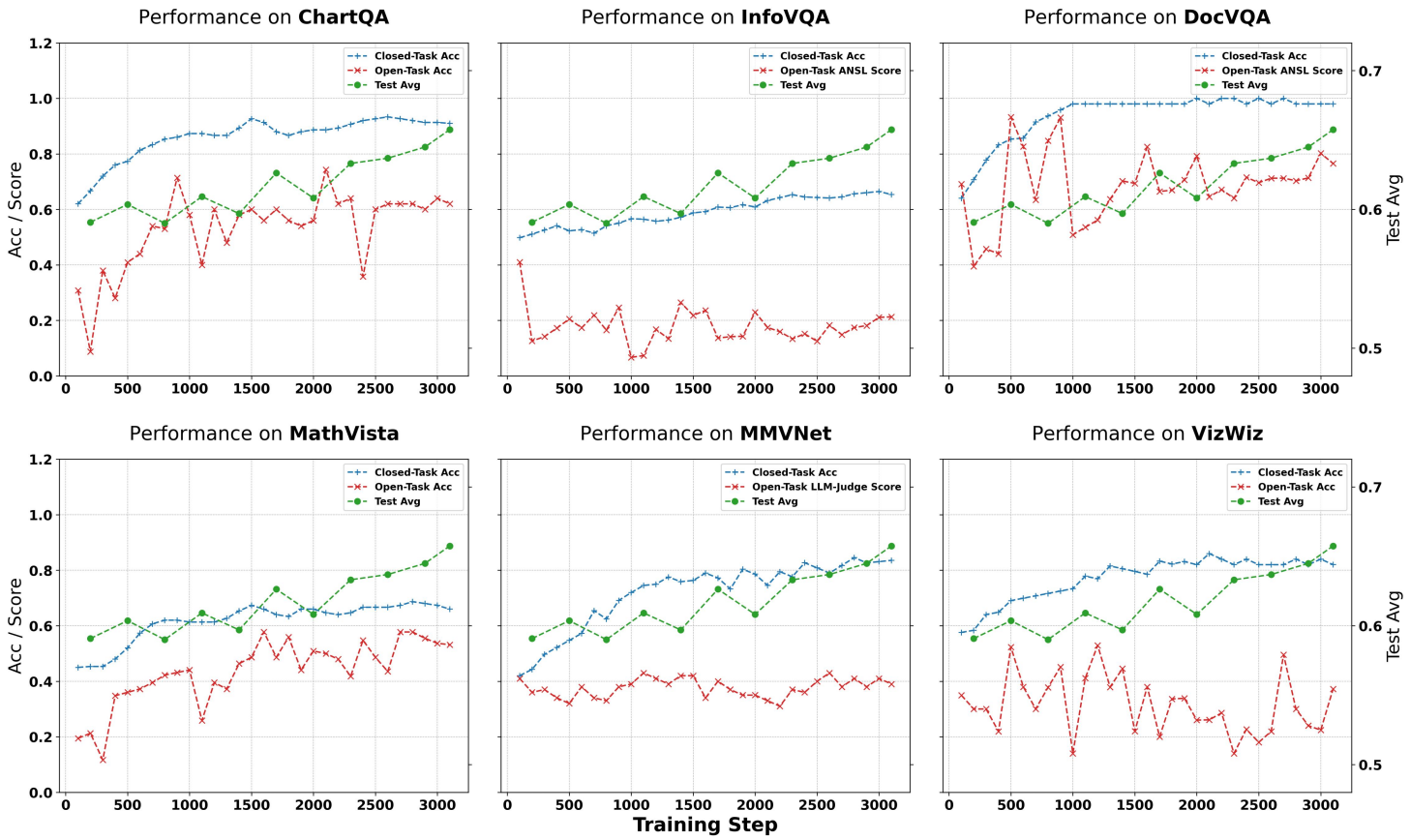


Figure 2: Validation Trajectories across 6 Datasets. Comparison of **Open-Task (Red)** and **Closed-Task (Blue)** signals against the **"Gold Standard" Test Average (Green)**. The Closed-Task signal effectively filters out generation noise, strictly tracking the model's true performance trend.

Visual Stability

- **Open-Task (Red)**: Extremely volatile and noisy.
- **Closed-Task (Blue)**: Smooth, monotonic improvement.

Tracking the Truth:

- The Blue curve closely mirrors the **"Gold Standard" (Green)** trajectory.
- The Red curve diverges significantly.

Statistical Proof:

- Open-Task: Negligible correlation **(r = 0.061 ↓).**
- Closed-Task: Strong positive correlation **(r = 0.798 ↑).**

Finding 2: Orders-of-Magnitude More Efficient

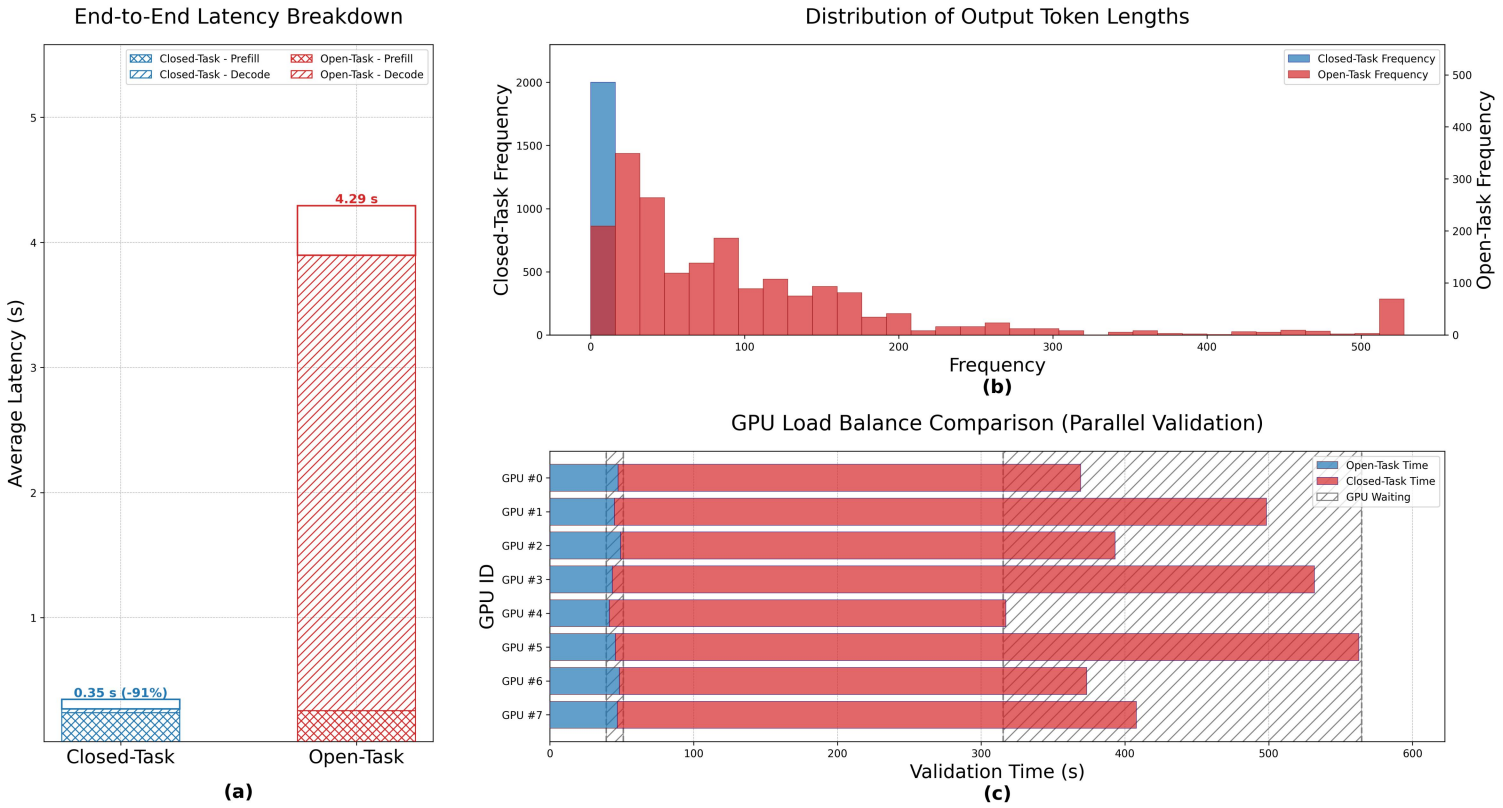


Figure 2: Efficiency Analysis. (a) A 91% latency reduction is achieved by removing the decode step. (b) Open-Task output lengths follow **a stochastic long-tail distribution (Red)**, whereas Closed-Task is **deterministic (Blue)**. (c) This long tail causes severe GPU load imbalance and waiting time (hatched area) for Open-Task, while Closed-Task achieves near-perfect parallelization.

>10x Speedup :

- **Open-Task (Red)**: Extremely volatile and noisy.
- **Closed-Task (Blue)**: Smooth, monotonic improvement.

Eliminating Stragglers :

- The Blue curve closely mirrors the **"Gold Standard" (Green)** trajectory.
- The Red curve diverges significantly.

Perfect Load Balance :

- Open-Task: Negligible correlation **(r = 0.061 ↓).**
- Closed-Task: Strong positive correlation **(r = 0.798 ↑).**

Conclusion: A Better Compass for VLM Training

The Problem :

- Standard Open-Ended validation is unreliable (r=0.061, high variance) and inefficient (straggler issues) .

The Solution :

- Closed-Task Validation (via Multiple-Choice conversion) bypasses the decoding bottleneck.

The Impact :

- Reliable: Provides a stable signal highly correlated with final performance (r=0.798).
- Efficient: Achieves >10x speedup with near-perfect system utilization.
- Verdict: A robust methodology bridging rapid iteration and scientific reliability.

Thank You

Q & A