Sparse-Checklist Prompting for Arabic Grammar Tutoring: Fast, Token-Efficient Feedback

Zayan Hasan
Rock Ridge High School — Ashburn, VA, USA
zayanuhasan@gmail.com
Muslims in ML (MusIML) @ NeurIPS 2025

Motivation

Muslim community classrooms often serve large groups of beginner Arabic learners with limited instructional time. Teachers need **fast, low-cost, targeted** feedback for independent practice.

Arabic morphology (agreement, clitics, definiteness) produces many small-but-important errors where **short, specific hints** help more than long explanations.

LLMs can support tutoring, but free-form reasoning is **slow, verbose, expensive**. We explore a **token-efficient, constrained-output** alternative.

Methods: Sparse-Checklist Prompting

We compare three prompting strategies:

- **Direct**: CORRECT/INCORRECT + short explanation (up to 80 completion tokens)
- **Sparse-Checklist**: label + one tag from a 5-skill taxonomy: AGREEMENT, PRONOUN_CLITIC, PREPOSITION, DEFINITENESS, NONE
- **Router**: if student's choice = gold, use short Direct; otherwise use Sparse-Checklist

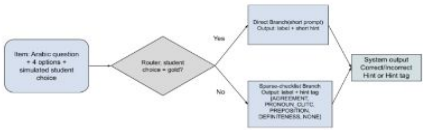

Figure 1: Sparse-Checklist prompting with routing.

Key idea: Treat explanation as **selection**. Selecting a tag produces interpretable hinting while keeping completion length short.

Experimental Setup

- **Dataset:** 180 synthetic MCQs (45 per skill)
- **Model:** GPT-4o-mini, temperature=0
- **Token caps:** Direct=80, Checklist=40, Router=40
- **Metrics:** accuracy, macro-F1, median latency, completion tokens ("reasoning cost"), hint-tag accuracy

Main Results

Method	Acc.	F1	Lat.	Tok.	Tag Acc.
Direct	76.1	76.1	0.807	22.7	0
Sparse-Checklist	**81.1**	**81.1**	**0.530**	**11.9**	**100**
Router	79.4	79.4	0.639	18.2	100

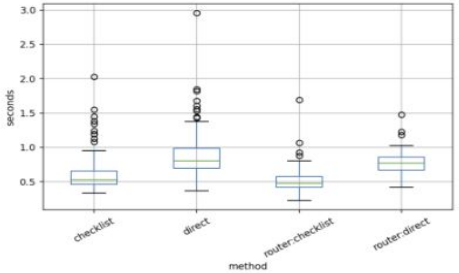

Figure 2: Latency distributions (boxplots).

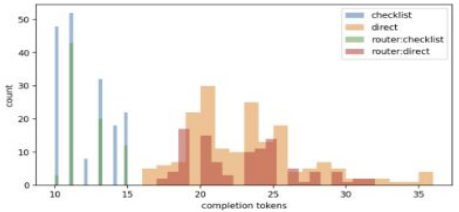

Figure 3: Completion token counts (histogram).

Key Findings

- **+5% correctness improvement** over Direct.
- Latency reduced from **0.807** → **0.530** seconds.
- Completion tokens nearly halved (**22.7** → **11.9**).
- Router provides a balanced cost–accuracy tradeoff.
- Sparse-Checklist achieves **100% correct tag selection** on incorrect items.

Per-Skill Performance

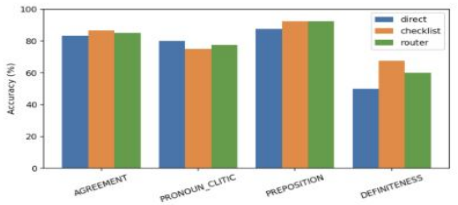

Figure 4: Accuracy by skill.

Consistent improvements across all four grammar categories:

- Agreement
- Pronoun Clitics
- Prepositions
- Definiteness

Limitations

- Synthetic-only evaluation.
- Only one model tested.
- Router uses oracle routing (upper bound).
- Single-tag hints may miss multi-skill errors.
- Dataset size small (180 items).

Conclusion

Sparse-Checklist converts LLM reasoning into **fast, token-efficient tag predictions**. Across 180 items, it improves correctness, reduces latency, and halves reasoning cost—while preserving perfect hint accuracy.

A lightweight, model-agnostic approach suited for Muslim community classrooms with limited resources.