30th ANNIVERSARY EMNLP 2025
Suzhou, China | 中国苏州
November 4-9 | 11月4日-9日

UNIVERSITY OF CAMBRIDGE

BLiSS 1.0: Evaluating Bilingual Learner Competence in Second Language Small Language Models

Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, Weiwei Sun

Department of Computer Science and Technology, University of Cambridge
ALTA Institute, University of Cambridge

This work is supported by
Cambridge University Press & Assessment

The Challenge in Evaluating Acquisition-oriented Language Models

- Current benchmarks evaluate models on native-like grammaticality.
- Human L2 learners, however, produce systematic, predictable "errors" (interlanguage).
- **The Gap:** No benchmarks test for cognitively plausible, human-like error patterns.

	Standard LMs Focus: Grammaticality	L2 Learners Process: Interlanguage
We play sports.	🟢	🟢
*We play **the** sports.*	🔴	🟢
*We **the** play sports.*	🔴	🔴

How can we test if a model prefers a plausible human error over a contrived one?

Our Paradigm: Selective Tolerance

- To solve the evaluation gap, we need a method that can measure two things at once: a model's grammatical knowledge AND its sensitivity to human-like error patterns. A simple "right vs. wrong" test can't do this.

Principle 1: Isolate Sensitivity to Plausibility.
We must test if the model can distinguish a systematic, human-like error from a contrived one

Principle 2: Ground the Test in Grammaticality
This sensitivity must not come at the cost of grammatical competence.

The BLiSS Benchmark

- Over 2.8 million raw learner sentences from large corpora (EFCAMDAT, W&I, FCE).
- Yields 136,867 high-quality triplets after a rigorous validation pipeline.
- Includes rich metadata: Learner L1, CEFR proficiency, and error type.
- **Key Unit: The Plausibility Triplet**

🟢 **Corrected**
There are a lot of benefits when we play sports.

🟡 **Naturalistic Learner Error**
There are a lot of benefits when we play **the** sports.

🔴 **Contrived Artificial Error**
There are a lot of benefits when **the** we play sports.


```
{
  "learnerID": "8421",
  "L1": "Vietnamese",
  "cefr": "C1",
  "topic": "play  sports ",
  "corrected": "There are a lot of
     benefits when we play  sports .",
  "learner error": "There are a lot of
      benefits when we play  the
     sports .",
  "artificial error": "There are a lot
      of benefits when the we play
     sports .",

  "errant_edits": [{
    "type": "U:DET",
    "o_str": "the",
    "c_str": ""
  }],
  "all_error_types": [
    "U:DET"
  ]
}
```


Quantifying Selective Tolerance

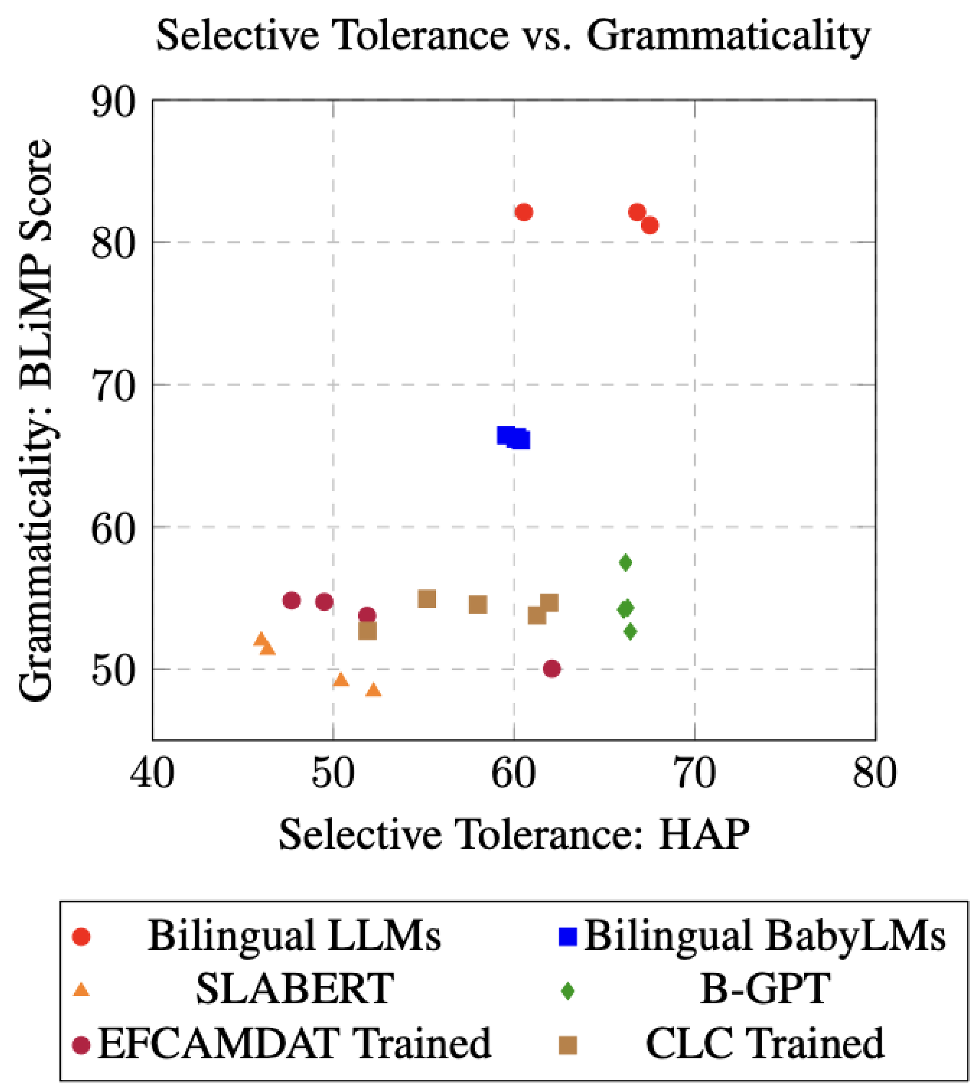
Scoring: Token-normalized Surprisal (BPT) measures plausibility.

Metrics:

- **HAP (Human vs. Artificial):**
  Does the model prefer 🟡 over 🔴?
  (Tests for basic selective tolerance)
- **SO (Strict Order):**
  Does the model rank 🟢 < 🟡 < 🔴?
  (The strictest test of grammar and tolerance)
- **LP (Learner Preference):**
  Does the model prefer 🟡 over 🟢?
  (A diagnostic for over-imitating errors)

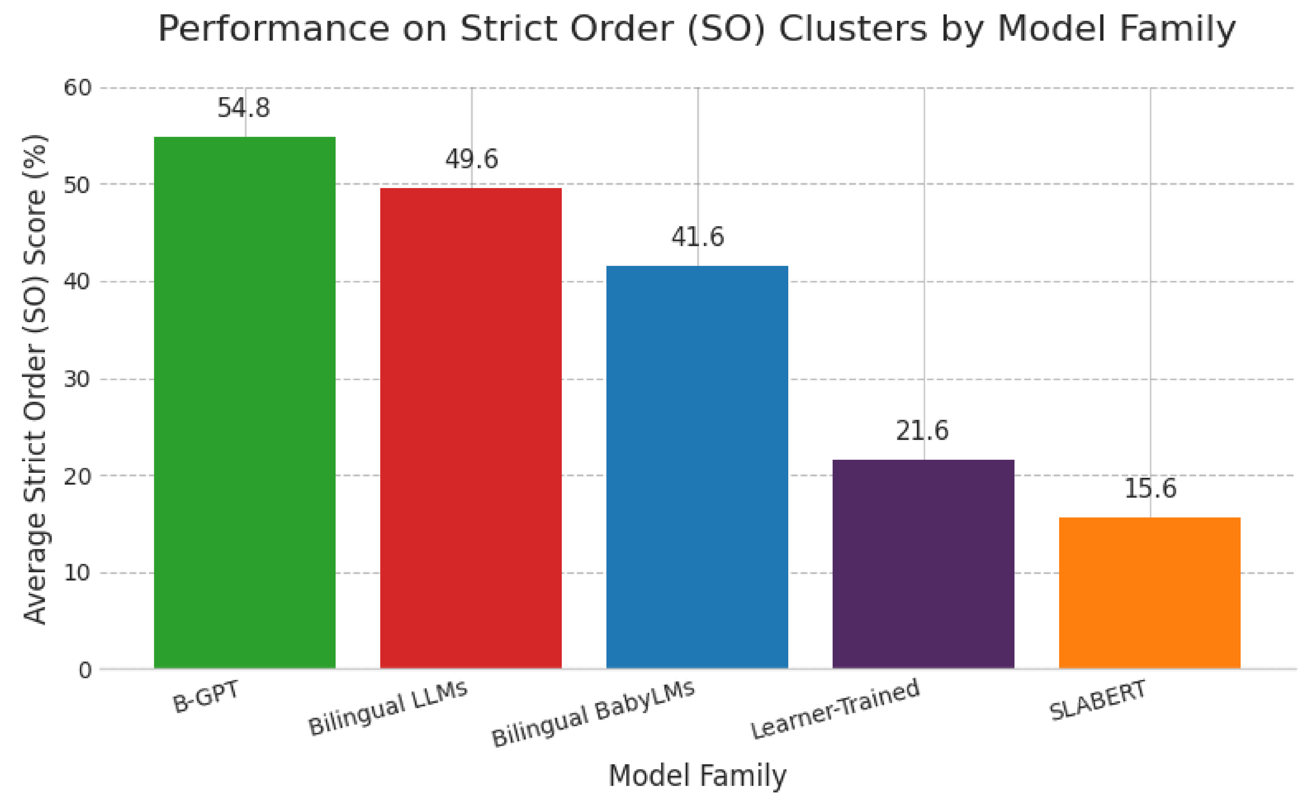
Key Findings

Finding 1: Selective Tolerance is Distinct from Grammaticality. High Grammaticality Does Not Guarantee High Selective Tolerance.



Takeaway: BLiSS measures a complementary skill. A model can master formal grammar but still fail to understand the nuanced patterns of human learner errors.

Finding 2: Training Paradigm is the Key Predictor of Performance.



Takeaway: A model's training paradigm is the strongest predictor of its performance.