GPT or BERT: Why not both? (Charpentier & Samuel, 2024)

Critique by Jingying Xu

November, 2025

The hybrid GPT-BERT model (Charpentier & Samuel, 2024) is the winner of the 2nd BabyLM Challenge (2024), a competition aimed at developing data-efficient pretraining methods using child-like amounts of data (10M–100M words) and providing insight into how humans successfully acquire language with limited linguistic input (Choshen et al., 2024). The challenge evaluates models on the BabyLM evaluation suite, which includes BLiMP/BLiMP-S for syntactic acceptability judgments, a GLUE subset for downstream understanding, EWoK for basic world knowledge plausibility, and LAMBADA for long-context prediction. Within this setting, the paper proposes a hybrid training objective that unifies BERT-style masked language modeling and GPT-style causal next-token prediction. The key idea is **Masked Next-Token Prediction**: during masked training, the model predicts token *k+1*at position *k*, which allows a single shared transformer to behave like BERT (bidirectional mask) or GPT (causal mask) without changing the architecture.

Training is conducted in two tracks: **STRICT-SMALL** (≈10M tokens, ≈30M parameters) and **STRICT** (≈100M tokens, ≈119M parameters), using a 1:1:1 data mix of BabyLM corpus, FineWeb-Edu, and Cosmopedia. The model also incorporates several sample-efficiency modifications (attention gating, layer-mixing, mask scheduling, batch-size ramping). Results show that GPT-BERT achieves top performance across reported systems: in STRICT-SMALL it reaches 81.2 BLiMP, 76.5 GLUE, and 54.6 EWoK, surpassing both the baseline LTG-BERT and last year's winner ELC-BERT; in STRICT it reaches 86.1 BLiMP, 81.5 GLUE, and 58.4 EWoK, again the strongest among available baselines. These results demonstrate that the hybrid objective **does improve data efficiency** in the BabyLM setting.

However, while the model succeeds at adult-style linguistic benchmarks, its contributions to the scientific study of human language acquisition are more limited. The goals of BabyLM and child language research differ: BabyLM optimizes task performance, whereas acquisition research seeks **explanatory adequacy**, specifically, to discover the mechanisms and constraints by which all humans acquire language. In child language acquisition, **mistakes are meaningful**: the specific errors children make (and the ones they never make) provide crucial evidence about the nature of their internal grammatical representations. For example, young children often omit auxiliaries (*Girl happy?*), but they never use the incorrect linear rule "move the first auxiliary" when forming yes-no questions (e.g., *The girl who **is** smiling **is** happy* → **Is the girl who smiling **is** happy?*). Instead, they apply the structure-dependent rule "move the auxiliary of the matrix clause" (e.g., *The girl who **is** smiling **is** happy* → ***Is** the girl who **is** smiling happy?*) from the start without exception (Crain & Nakayama, 1987). In this sense, errors are not noise; they are diagnostic data about the grammar being learned. By contrast, BabyLM evaluation rewards correctness, not the shape of errors. High BabyLM scores therefore do not imply that the model learns human-like grammatical representations or developmental constraints. High accuracy on BLiMP or GLUE shows task success, but does not reveal what kinds of grammatical generalizations the model is making. In relation to *Stochastic Parrots* (Bender et al., 2021), the model has the advantage of small-scale, controlled data, but the paper does not address whether the model truly understands or acquires language in a human-like way. Overall, GPT-BERT is a strong methodological contribution to efficient pretraining, but not a theory of how children acquire language.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *FAccT*.

Charpentier, G. G., & Samuel, D. (2024). GPT or BERT: Why not both? *Proceedings of the 2024 Conference on Computational Natural Language Learning*.

Choshen, L., et al. (2024). The BabyLM Challenge 2024: Sample-efficient pretraining on a developmentally plausible corpus. *BabyLM Workshop*.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63*(3), 522–543.