Pretraining Objective Choice in Multilingual Data-Constrained Shared Encoders

Anonymous ACL submission

Abstract

Which pretraining objective should we use when data is scarce? Prior comparisons fix scale, language, or evaluation paradigm, leaving the joint choice open in the data-constrained regime. We compare seven shared-encoder objectives across four languages (English, Dutch, Chinese, Korean) at 10M and 100M words under a matched token budget, optimizer, and schedule. No objective is uniformly best: CLM is strict-best on BLiMP-family grammaticality in all four languages, an MLM-led bidirectional cluster leads English GLUE at 10M and narrows to within noise by 100M, and the bidirectional-vs-AR ordering inverts in Korean, the single agglutinative language in our sweep, at 30M words (+3.34pp for the bidirectional cluster). On the same encoder, BLiMP and GLUE rankings correlate weakly across the seven objectives, and zero-shot commonsense does not separate objectives at 30–120M parameters. Objective choice in the data-constrained regime is conditional on the evaluation paradigm, language, and scale. Our results indicate that single-language sweeps under-determine objective choice for the multilingual sub-100M-word regime. Our code, evaluation harness, and analysis scripts are available at https://anonymous.4open.science/r/objective-choice-data-constrained-B95A/README.md.

1 Introduction

The choice of pretraining objective is largely settled at scale. GPT-style decoder-only language models use causal language modeling (CLM, i.e. next-token prediction) (Radford et al., 2018, 2019; Brown et al., 2020), and BERT-style encoders use masked language modeling (MLM, i.e. predicting masked tokens from both-side context) (Devlin et al., 2019). These two families rarely compete at the same scale, and the choice remains open in the *data-constrained* regime (Warstadt et al., 2023) and in many non-English settings. *Which objective wins when data is the binding constraint?*

Prior work fixes one axis at a time. Raffel et al. (2020) compare objectives at C4 scale; Muennighoff et al. (2023) characterize the cost of data repetition under fixed CLM; in-regime studies optimize pipelines over a single objective (Samuel et al., 2023; Charpentier and Samuel, 2024); Artetxe et al. (2022) show that the application-conditional bidirectional advantage (helpful for fine-tuning and infilling, harmful for next-token prediction) holds up to 6.7B parameters in English. This paper compares seven objectives (Devlin et al., 2019; Radford et al., 2018; Charpentier and Samuel, 2024; Raffel et al., 2020; Dong et al., 2019; Edman and Fraser, 2025; Clark et al., 2020) on English, Dutch, Chinese, and Korean at 10M and 100M words with 30M midpoints, for a total of 269 pretraining runs. Three pretrain seeds are used in every multi-seed cell, and three fine-tuning seeds are added at English 100M and at Dutch, Chinese, and Korean 30M and 100M, restricted to the five-objective "5-core" subset defined in §3. The scope is the data-constrained regime (Warstadt et al., 2023; Ahia et al., 2023; Eldan and Li, 2023); the $\geq$1B-parameter regime is out of scope. Four findings, labeled R1 through R4, are fixed on English 10M before any 100M result is analyzed; this is an internal held-out protocol rather than an external preregistration (Appendix A.15). The other cells then serve as held-out transfer checks across languages and scales (§3, Table 1).

Contributions. (i) **A matched-budget multilingual sweep.** This study compares seven shared-encoder objectives across four typologically distinct languages (English, Dutch, Chinese, Korean). Each (language, scale) cell trains a monolingual model from scratch, with no joint multilingual pretraining. Encoder–decoder denoising methods such as T5, BART, and UL2 vary input–output factor-

ization as well as objective; these are treated as a separate axis and left to future work.

(ii) **A cross-language trajectory with a single sign-flipped cell.** On English GLUE at 10M words, an MLM-led bidirectional cluster leads CLM: the MLM–CLM contrast survives Holm at family-wise α=0.05, while AMLM-soft and Hybrid are Holm-borderline. By 100M words, the five core objectives collapse into the heuristic floor and become mutually indistinguishable. Dutch follows the same narrowing trajectory; Chinese is at-noise from 10M onward. Korean inverts the pattern at the 30M midpoint, with a sign-flipped bidirectional advantage on `klue_nli` and `klue_re` (`klue_sts` goes the other way). The reversal is corpus-controlled by the shared KO 10M/30M source and robust across nine paired KLUE trials, and is consistent with a Korean-specific, morphology-sensitive hypothesis.

(iii) **A per-cell summary of the best-observed family.** Table 1 reports the top-tier family for each (target, language, scale) cell. The summary is needed because within-cell rankings do not transfer cleanly: same-encoder BLiMP–GLUE rank correlation is a weak point estimate (§5), and no single objective wins across evaluation targets under our predefined top-tier accounting (Appendix A.12).

(iv) **A verified negative scope within $\leq$120M parameters.** The likelihood-scoring alignment of Salazar et al. (2020) extends from English to Dutch, Chinese, and Korean and to two variants the original PLL work does not cover: AMLM-soft and ELECTRA's embedding-tied-generator PLL. Commonsense reasoning at 30–120M parameters stays near chance with no consistent winner: worst-to-best per-language gaps fall within 2$\times$ the binomial 95% CI of each other on every cell (Appendix A.14). This is a pre-registered prediction (R4 in §3.2) that survived held-out testing across four languages and a 120M-parameter capacity probe, bounding the reach of contributions (ii) and (iii) within the tested scale. Phenomenon-level notes are in Appendices A.18 and A.22.

2 Related Work

Objective comparisons at scale. Raffel et al. (2020) compared denoising objectives at C4 scale; Dong et al. (2019) unify bidirectional, unidirectional, and sequence-to-sequence objectives in a single model; Lewis et al. (2020) introduce denoising sequence-to-sequence pretraining; Joshi et al. (2020) predict span boundaries without sentinel tokens (we adopt the T5-style sentinel formulation for our Span row); Clark et al. (2020) introduce replaced-token detection; Tay et al. (2023) mix several denoising styles at scale; and Aribandi et al. (2022) layer extreme multi-task mixtures on top of span corruption. The closest small-encoder objective sweep is Aroca-Ouellette and Rudzicz (2020), who compare 14 *auxiliary* sentence-level losses (next-sentence-prediction variants, sentence ordering, and term-frequency prediction) on BERT-style encoders in English; they fix the primary MLM objective and vary only the auxiliary signal. We instead vary the primary token-level objective itself across seven choices and four languages, and our 5-core / non-core split is orthogonal to their auxiliary-loss axis. At $\geq$1B parameters, Wang et al. (2022) report that the best objective–architecture pairing depends on the application, and Artetxe et al. (2022) show that the application-conditional bidirectional pattern (helpful for fine-tuning and infilling, harmful for next-token prediction) is preserved up to 6.7B parameters within task families. We re-examine these patterns in the extreme small-data regime. The English GLUE bidirectional advantage is pronounced at 10M words and collapses into the bootstrap noise floor by 100M, while Korean at 30M departs from this trajectory and Dutch and Chinese are consistent with it (§5).

Data-constrained pretraining. The 10M–100M word regime has been used as a controlled scarcity setting for studying small-encoder pretraining (Warstadt et al., 2023; Charpentier et al., 2025), and the present study adopts this regime without making any acquisition-mechanism claim. Existing in-regime studies usually evaluate a single tuned pretraining recipe at a time. Samuel et al. (2023) train MLM on the 100M-word British National Corpus and propose LTG-BERT; Charpentier and Samuel (2024) share CLM and MLM heads on one encoder (GPT-BERT, the architecture adopted here); Yu et al. (2024) alternate causal and masked phases. Muennighoff et al. (2023) quantify the loss penalty from repeating training data under fixed CLM, finding up to 4 epochs of repetition incurs negligible loss. The orthogonal axis is the one targeted here: *at matched schedule and token budget, which objective?*

Multilingual low-resource pretraining. Conneau and Lample (2019); Conneau et al. (2020); Xue et al. (2021) scale masked and translation language modeling and span corruption to multilin-

gual data, but they vary scale or task and fix the objective per language. Ahia et al. (2023) document the cost of tokenizer mismatch under data scarcity; Eldan and Li (2023) show that small models can be competent below production scale; and Baby-BabelLM (Jumelet et al., 2026a) extends matched-budget pretraining to multiple languages. In the $\leq$100M-word range, objective choice is rarely re-examined per language. The current study evaluates seven objectives on four languages spanning the Germanic family (English, Dutch), Sinitic (Chinese), and Koreanic (Korean), with language-appropriate BLiMP-family minimal-pair benchmarks, GLUE-family fine-tuning suites, and commonsense benchmarks (§4.4).

Adaptive masking and scaling laws. Edman and Fraser (2025) introduce adaptive masked language modeling (AMLM-hard/-soft) that modulates per-token mask probabilities by token difficulty; the soft variant is used here. Levine et al. (2021) mask correlated n-grams jointly via PMI; that work uses correlation structure, not per-token difficulty. Kaplan et al. (2020); Hoffmann et al. (2022); Brown et al. (2020) characterize loss scaling under fixed AR objective in data-abundant settings; at matched schedule and budget the *shape* of the scaling curve is objective-dependent (§5.3, Appendix A.11).

Evaluation paradigms. Warstadt et al. (2020) aggregate 67 minimal-pair paradigms into the BLiMP score; our subcategory analysis (Appendix A.22) shows that this aggregation hides phenomenon-level differences between objectives. Salazar et al. (2020) introduce pseudo-log-likelihood (PLL) for bidirectional models, since these models do not produce a native autoregressive likelihood. The resulting training-versus-scoring mismatch is a structural cause of ranking divergence between BLiMP and PPL.

Positioning. No prior study compares seven objectives across four typologically distinct languages in the $\leq$100M-word regime under matched schedule and budget. The closest objective sweeps (Raffel et al., 2020; Tay et al., 2023; Wang et al., 2022) are EN-only at $\geq$1B tokens, and multilingual pretraining (Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021) fixes the objective per study. Per-step FLOPs vary by objective and step counts vary by language token budget (Appendix Table 7). Scale is modest ($\leq$120M parameters); systematic coverage enables the framework that follows.

3 Setup and Empirical Predictions

The best pretraining objective depends on what is measured. The analysis is organized around four empirical regularities, labeled R1 through R4 and defined in §3.2: R1 covers likelihood-mode alignment (which objective tops which scoring mode), R2 covers grammatical acceptability (BLiMP-family), R3 covers downstream transfer (GLUE-family fine-tuning), and R4 covers commonsense reasoning at 30–120M parameters. R1 and R2 hold uniformly across all four languages in the sweep. R3 varies with scale, holding strongly at English 10M and collapsing into the noise floor by 100M, with Dutch and Chinese consistent with this trajectory and Korean 30M inverting the bidirectional pattern. R4 is a near-chance null. Table 1 compiles the observed top-tier family per condition. The patterns are checked as held-out transfer tests across Dutch, Chinese, and Korean at the 100M scale before the summary table is drawn; a Korean boundary-feature probe is reported in §5 and Appendix A.20.

Terminology. The *5-core* comprises CLM, MLM, Hybrid, Span, and AMLM-soft; ELECTRA and Prefix LM sit outside this group, each diverging from the 5-core under direct bootstrap CIs on EN 100M GLUE (Table 10). The *bidirectional infill* group covers MLM, Span, and AMLM-soft (token prediction with both-side context); Hybrid and Prefix LM are mixed forms (CLM-mode on disjoint batch halves; CLM-mode after a bidirectional prefix). *Scoring modes* are autoregressive (AR; CLM, Hybrid, Prefix LM), pseudo-log-likelihood (PLL; MLM, Span, AMLM-soft) (Salazar et al., 2020), and embedding-tied-generator PLL (ELECTRA); rankings are within a mode unless stated.

Table 1: Observed top family per condition for the 30M-parameter encoder at 10M and 100M data. Rule labels R1 through R4 are defined and discussed in §3.2; detailed statistics appear in §5 and the 120M-parameter P6 probe in Appendix A.11.

Condition	Top family	Rule
PPL (AR mode)	CLM	R1
PPL (PLL mode)	AMLM-soft	R1
BLiMP, 100M	CLM in all 4 languages	R2
EN GLUE, 10M	Bidirectional cluster	R3
EN GLUE, 100M	5-core within floor	R3
Cross-lingual GLUE	KO 30M inverts	R3
Commonsense, 30–120M	Not discriminative	R4

3.1 Characterizing Pretraining Objectives

Table 2 catalogs each objective along five properties; we use these as a labeling convention, not a predictive framework. The properties are *Flow* (left-to-right, both directions, or mixed), *Frac.* (loss-contributing positions per step), *Mult.* (loss heads; two for Hybrid (Charpentier and Samuel, 2024)), *Adapt.* (learned masking, AMLM-soft only (Edman and Fraser, 2025)), and *Type* (generative; discriminative for ELECTRA).

Table 2: Seven objectives along five properties. *Frac.* is the per-head fraction of positions contributing to the loss (Hybrid uses 100%/15% on disjoint batch halves). *Mult.* is the number of independent loss heads. *Adapt.* marks objectives with learned masking. *Type* is generative or discriminative.

Objective	Flow	Frac.	Mult.	Adapt.	Type
CLM	uni.	100%	1	no	gen.
MLM	bi.	15%	1	no	gen.
Hybrid	mixed	100%/15%	2	no	gen.
Span	bi.	15%	1	no	gen.
Prefix LM	mixed	∼50%	1	no	gen.
AMLM-soft	bi.	15%	1	yes	gen.
ELECTRA	bi.	100%	1	no	disc.

3.2 Three Empirical Regularities and One Negative Scope

Table 1 is supported by three empirical regularities (R1, R2, R3) and a separately reported negative scope finding (R4). Each is stated as an a priori expectation grounded in the scoring or training protocol, and then tested on held-out cells.

R1 (cross-language likelihood alignment). A calibration check. Across EN/NL/ZH/KO, AR objectives top AR likelihood and bidirectional infill top PLL, extending Salazar et al. (2020) from English to four languages and to two variants outside that work (AMLM-soft; ELECTRA embedding-tied-generator PLL).

R2 (grammatical acceptability). Minimal-pair benchmarks reward calibrated token probabilities. AR training estimates the joint sequence probability end-to-end; bidirectional infill sees both-side context but does not commit at training time. CLM is therefore expected to lead the BLiMP-family mean, with bidirectional objectives within the floor. The expectation holds at 100M in all four languages; at smaller scales KBLiMP departs (Hybrid at 10M, AMLM-soft at 30M; CLM within the floor of the leader at both, Appendix A.8). Appendix A.6 reports the scorer-sensitivity check.

R3 (English-specific, regime-bounded). Bidirectional context provides richer per-position features at low data, and AR accumulates this signal more slowly as data grows. A bidirectional EN GLUE advantage at 10M, narrowing toward 100M, is therefore expected. At EN 10M, an MLM-led bidirectional cluster leads CLM; MLM–CLM survives Holm at α=0.05, with AMLM-soft–CLM and Hybrid–CLM Holm-borderline (Appendix A.5). By EN 100M the five core objectives are no longer separated under our bootstrap CIs or the 0.83pp floor. R3 is a transfer hypothesis on NL/ZH/KO: the direction holds at NL 10M→100M, ZH is at-noise from 10M, and KO 30M inverts (§5).

R4 (negative scope on commonsense). The three commonsense tasks (§4.4) stay near chance at 30M parameters with no consistent cross-language winner. The claim is restricted to 30–120M parameters. R4 would be falsified if a single objective led on ≥2 of 3 commonsense tasks in ≥3 of 4 languages with a per-task gap above 2pp.

R2 and R3 are fixed on EN 10M before any 100M result is consulted (internal held-out protocol; not an external preregistration, Appendix A.15); NL/ZH/KO 100M serve as out-of-distribution tests. The 1.96σ heuristic floor (2.6pp BLiMP, 1.56pp GLUE on the 3-seed pool) is a tier convention; significance claims rest on direct bootstrap CIs.

4 Experimental Setup

4.1 Pretraining Objectives

Seven pretraining objectives are compared (Table 2; hyperparameters in Appendix A.2): **MLM** (Devlin et al., 2019) (15% masked, 80/10/10); **CLM** (Radford et al., 2018) (left-context next-token); **Hybrid** (CLM+MLM) (Charpentier and Samuel, 2024) (two heads on a shared encoder, losses summed on disjoint batch halves); **Span corruption** (Raffel et al., 2020) (T5-style sentinels, mean span ≈ 3); **Prefix LM** (Dong et al., 2019; Raffel et al., 2020) (bidirectional prefix, autoregressive suffix); **AMLM-soft** (Edman and Fraser, 2025) (loss-modulated per-token mask probability); **ELECTRA** (Clark et al., 2020) (generator-discriminator replaced-token detection). The 5-

core / non-core split follows the empirical definition in §3.

4.2 Data

The study covers 10M and 100M words per language, plus a 30M midpoint that traces the trajectory. **10M**: the BabyLM 10M corpus (Warstadt et al., 2023; Charpentier et al., 2025) for EN, and BabyBabelLM (Jumelet et al., 2026a) matched to 10M for NL/ZH/KO. **30M**: NL/ZH/KO are strict supersets of the 10M sample (same source corpus, controlling for corpus effects); EN 30M (P4.c) uses a CHILDES-heavy 30M English subsample (EN 30M PPL excluded from cross-scale comparison). The CC-100 EN 10M cell (P5) is a domain check. **100M**: the 100M EN corpus, the BabyBabelLM 100M splits for NL/ZH (ZH sized by a 1.01B-character budget), and CC-100 for KO (no 100M BabyBabelLM Korean split; discussed in §6). Token counts and epoch settings in Appendix Table 7 (139M–1.01B tokens, 2 or 10 epochs). One byte-level BPE (Sennrich et al., 2016; Radford et al., 2019) per language at 16K vocabulary, trained on 10M and reused at 30M and 100M; the Ġ prefix marks word-initial tokens (used by the Korean boundary probe, Appendix A.20).

4.3 Model and Training

All experiments use GPT-BERT (Charpentier and Samuel, 2024), which supports both causal and masked objectives on a single model by switching the attention mask. The default has about 30M non-embedding parameters across 12 layers, hidden size 384, 6 attention heads, and feed-forward width 1536. A 120M-parameter variant used in P6 doubles both depth and width at EN 10M.

Training uses AdamW (Loshchilov and Hutter, 2019) at peak LR 10^{-3}, 2% warmup with cosine decay, bf16, batch 128, sequence length 128, on a single H100. Within each (language, scale) cell, all seven objectives share optimizer, schedule, batch, sequence length, and step count. Token budget is matched per cell; wall-clock FLOPs are not (Hybrid ≈1.3× and ELECTRA ≈1.5× the CLM per-step cost). All cells train for 10 epochs except ZH 100M (2 epochs, matching the character-token budget). Pretrain seeds are 42, 123, 456 in every multi-seed cell. GLUE FT (Table 3) uses 9 trials (3 pretrain × 3 FT) at EN 100M (all 7 obj) and at NL/ZH/KO 30M/100M (5-core only); ELECTRA, Prefix LM, and EN 30M use FT 42 alone (3 trials; SDs reflect pretrain-seed variance).

Table 3: GLUE FT seed protocol. 9-trial cells: EN 100M (all 7 obj) and NL/ZH/KO 30M/100M (5-core). ELECTRA/Prefix LM and EN 30M use FT 42 (3 trials). Cross-lingual: NL DUMB (de Vries et al., 2023), ZH CLUE (Xu et al., 2020), KO KLUE (Park et al., 2021).

Cell	Pretrain	FT	n
EN 10M, 7 obj	3	1	3
EN 100M, 7 obj	3	3	**9**
NL/ZH/KO 30M, 5 core (P4)	3	3	**9**
NL/ZH/KO 30M, ELECTRA/Prefix	3	1	3
NL/ZH/KO 100M, 5 core (P3)	3	3	**9**
NL/ZH/KO 100M, ELECTRA/Prefix	3	1	3
EN 30M, 5-core (P4.c)	3	1	3

4.4 Evaluation

Suite. Four evaluation families, with scoring matched to each objective. (i) **PPL**: AR likelihood for AR objectives, PLL (Salazar et al., 2020) for bidirectional infill, generator PLL for ELECTRA via the embedding-tied generator (Clark et al., 2020 tie embeddings only); ranks within mode and not compared across modes (incommensurable; Appendix A.24). (ii) **Minimal-pair acceptability**: BLiMP-family per language (BLiMP (Warstadt et al., 2020), BLiMP-NL (Suijkerbuijk et al., 2025), ZhoBLiMP (Liu et al., 2024), KBLiMP (Song et al., 2025)) plus MultiBLiMP (Jumelet et al., 2026b); ranks grammatical above ungrammatical sentences via likelihood, scored through lm-evaluation-harness (Gao et al., 2024). (iii) **Multiple-choice and coreference** (zero-shot): HellaSwag (Zellers et al., 2019) and XStoryCloze (Lin et al., 2022) score candidate continuations by likelihood; Winogrande (Sakaguchi et al., 2020) uses the same protocol on coreference candidates. HellaSwag and Winogrande are EN-only; every model is scored on them to document near-chance behavior, with no cross-lingual claim. (iv) **Downstream FT**: GLUE-family tasks (Wang et al., 2019b) fine-tune a task head on top of encoder representations, so the protocol is paradigm-independent. EN uses the 7-task GLUE+SuperGLUE mean (Wang et al., 2019b,a); NL/ZH/KO use DUMB (de Vries et al., 2023), CLUE (Xu et al., 2020), and KLUE (Park et al., 2021). Per-task lists: Appendix A.3.

Statistics. The 1.96σ floor (2.6pp BLiMP, 1.56pp GLUE 10M, 0.83pp GLUE 100M) is a tier convention, not an equivalence test. At n=3 a Student-t correction ($t_{2,0.025}$=4.30) widens the BLiMP and GLUE-10M bands to ∼5.7pp and ∼3.4pp; at n=9 the GLUE-100M band is essentially unchanged ($t_{8,0.025}$=2.31, ∼0.94pp). Load-

bearing claims rest on direct bootstrap 95% CIs (Efron and Tibshirani, 1994) with $N{=}10^4$ resamples and Holm-Bonferroni at family-wise $\alpha{=}0.05$ for multi-seed pairwise families, not on the floor; rule-critical contrasts must survive Holm. Per-cell details: Appendix A.5.

4.5 Experiment Matrix

Experiments are organized into six phases by their role in the argument (Appendix Table 8): **P1** EN 10M baseline where R3 is constructed (21 runs); **P2** cross-lingual 10M held-out for R1/R2 (63); **P3** 4-lang $\times$ 100M main grid for R1/R2/R3 endpoints (84, five sub-batches P3.a–P3.e); **P4** 30M trajectory midpoint (78; NL/ZH 7-obj 3-seed P4.a, KO 7-obj 3-seed P4.b, EN 5-core 3-seed P4.c); **P5** EN 10M CC-100 domain check (7); **P6** EN 10M 120M-parameter scaling probe (19; 5-core + ELECTRA at 3 seeds, plus an ELECTRA seed-789 rerun replacing the seed-42 generator-PLL outlier). Total: 269 pretraining runs on a single H100 GPU.

5 Results

Evidence for the four regularities is presented in order, R1 through R4 (definitions in §3.2, summary in Table 1 and Figure 1). The seed protocol is uniform across the matched grid: every 100M cell uses three pretrain seeds, EN 100M adds three fine-tuning seeds for nine trials per objective, and all 30M cells (NL, ZH, KO) use three pretrain seeds at fine-tuning seed 42. A difference is called *within the floor* when it falls inside the column-specific empirical 1.96σ band (2.6pp on BLiMP and 1.56pp on GLUE for the EN 10M 3-seed pool; 0.83pp on GLUE for the EN 100M 9-trial pool); this floor is an operational tier and not an equivalence test. Cells not separated from the leader under the floor are marked with $*$.

5.1 R1: Cross-lingual Likelihood Alignment

R1 concerns the likelihood axis. In all four languages, the autoregressive objectives top autoregressive likelihood and the bidirectional infill objectives top pseudo-log-likelihood (Table 4). Absolute perplexity is not comparable across languages because tokenizer fertility differs; Korean PPL is roughly 30 to 50 times English PPL (Appendix Table 22). The within-mode ranking, however, is preserved. The two scoring modes define separate tiers (Salazar et al., 2020), and the present study extends the original English-only result to four languages and to two scoring variants the original work does not cover: AMLM-soft, and the ELECTRA generator PLL routed through the embedding-tied generator (Appendix A.24).

Table 4: 100M held-out PPL within scoring mode (lower better). Blocks (PLL / AR / generator-PLL) are not comparable. 3-seed mean (SD).

Objective	EN	NL	ZH	KO
PLL scoring (bidirectional infill objectives)				
AMLM-soft	**5.59** (0.02)	**5.06** (0.07)	**10.7** (0.1)	**173.7** (0.9)
MLM	6.65 (0.20)	5.81 (0.11)	14.8 (0.3)	253.8 (24)
Span	8.33 (0.47)	7.02 (0.33)	27.4 (0.6)	575.4 (58)
AR scoring (autoregressive objectives)				
CLM	**25.75** (0.90)	**21.3** (1.0)	**111.2** (2.5)	**1,281** (14)
Hybrid	36.37 (0.28)	28.7 (0.4)	184.8 (6.9)	1,565 (24)
Prefix LM	305.0 (43.4)	187.2 (16.5)	391.6 (152.0)	4,611.6 (444.4)
Embedding-tied generator PLL (ELECTRA, no native likelihood)				
ELECTRA	10,330 (1628)	12,194 (239)	15,517 (946)	15,452 (382)

5.2 R2: Grammatical Acceptability (BLiMP)

CLM has the highest three-seed BLiMP mean in all four 100M languages (Table 5). Span is the runner-up on ZH and KO and third-best on EN and NL; the remaining objectives sit further below, and ELECTRA is last in every language. At EN 10M, 14 of 21 pairwise CIs survive Holm at $\alpha{=}0.05$ (Appendix A.5).

ELECTRA's last place on BLiMP reflects the scoring paradigm: ELECTRA is scored through the embedding-tied generator PLL, which is structurally incommensurable with the native AR and PLL columns, and a discriminator-only control (Appendix A.25) yields BLiMP scores indistinguishable from full ELECTRA. The BLiMP claim therefore applies to the shared-encoder, likelihood-scored protocol of this study. ELECTRA's last place on fine-tuned GLUE is paradigm-independent (representation-only) and remains a shared-encoder consideration (see Limitations).

Table 5: 100M acceptability accuracy (EN BLiMP, NL BLiMP-NL, ZH ZhoBLiMP, KO KBLiMP), 3-seed mean (SD). **Bold** = per-language top. Worst-to-best gap (CLM−ELECTRA): EN 9.6, NL 15.5, ZH 10.8, KO 16.1pp.

Objective	EN	NL	ZH	KO
CLM	**0.620** (0.022)	**0.648** (0.023)	**0.653** (0.007)	**0.628** (0.018)
Prefix LM	0.586 (0.007)	0.611 (0.017)	0.594 (0.016)	0.536 (0.027)
AMLM-soft	0.567 (0.011)	0.567 (0.013)	0.602 (0.020)	0.553 (0.005)
Span	0.571 (0.006)	0.617 (0.017)	0.628 (0.003)	0.605 (0.006)
Hybrid	0.566 (0.006)	0.565 (0.020)	0.590 (0.010)	0.569 (0.016)
MLM	0.566 (0.004)	0.588 (0.018)	0.625 (0.009)	0.594 (0.024)
ELECTRA	0.524 (0.022)	0.493 (0.044)	0.545 (0.009)	0.467 (0.012)

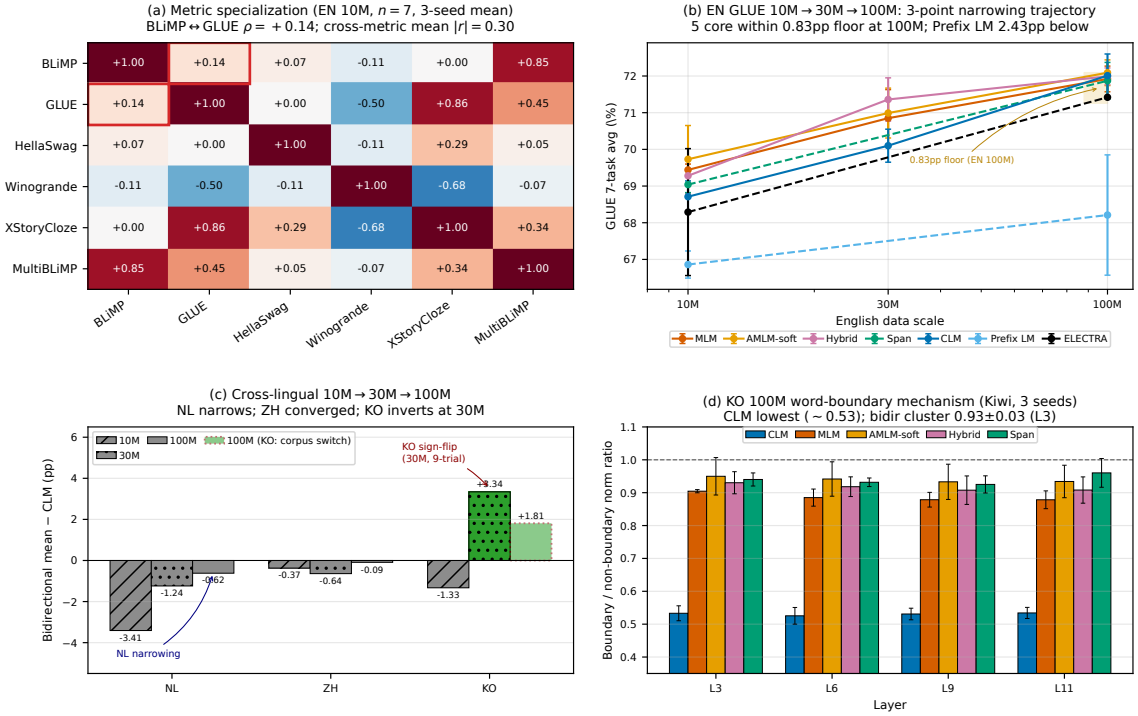

Figure 1: **Paper at a glance.** (a) At English 10M, BLiMP picks CLM and GLUE picks AMLM-soft, illustrating metric divergence (mean absolute pairwise Spearman $|\rho|$=0.30 over 15 pairs). (b) On English GLUE, the bidirectional advantage narrows from 10M to 100M, leaving the five core objectives within the floor at 100M. (c) On Dutch, Chinese, and Korean GLUE, Dutch narrows toward zero, Chinese is at-noise, and Korean inverts at the 30M midpoint with a bidirectional advantage of +3.34pp; the KO 100M cell uses CC-100 and is marked with a dashed border. (d) The Korean boundary-norm probe on 100M models (Kiwi tokenization, three seeds) separates CLM at 0.53 from the bidirectional cluster at 0.93 in layer 3.

5.3 R3: Downstream Transfer (English GLUE)

R3 holds at EN 10M and collapses by 100M. At 10M, an MLM-led bidirectional cluster leads CLM: MLM–CLM survives Holm at family-wise α=0.05, and AMLM-soft–CLM and Hybrid–CLM are Holm-borderline (adjusted thresholds between α=0.05 and α=0.1). CLM, Span, and ELECTRA sit 0.4–1.4pp below the leader and are not separated from one another at n=3 (Appendix A.5).

At 100M the separations collapse (Table 6): the five core are mutually indistinguishable (10/10 CIs include zero), Prefix LM sits 2.43pp below AMLM-soft, and ELECTRA falls within the floor but its pairwise CI excludes zero against each five-core objective (Appendix A.5). EN 30M anchors the trajectory (Appendix A.18); the 10M→100M gain is largest for CLM and smallest for the bidirectional cluster (Appendix A.23).

Cross-lingual GLUE 10M→30M→100M (Table 20). The 10M cells use three pretrain seeds at FT 42, and the 30M and 100M cells use 3 pretrain × 3 FT (9 trials per five-core cell). NL and ZH are consistent with the EN trajectory; KO 30M departs from it. NL narrows monotonically toward zero (30M and 100M within the floor of zero), and ZH is at-noise from 10M with all scale-to-scale gaps inside the floor. ZH 100M uses 2 epochs on a 1.01B-character budget, matching per-document

Table 6: EN downstream 7-task average (GLUE + SuperGLUE, mean±SD). 10M: 3 pretrain seeds, FT 42. 100M: 9 trials (3 pretrain × 3 FT). ⋆ = 5-core 100M cluster (CI vs. CLM includes zero); ICC + n_{eff} in Appendix A.5.

Objective	10M (3 pretrain seeds)	100M (n=9)
AMLM-soft	**69.73**±0.92	**72.09**±0.35⋆
CLM	68.71±0.43	72.01±0.59⋆
Hybrid	69.28±0.47	71.99±0.23⋆
MLM	69.44±0.16	71.92±0.35⋆
Span	69.04±0.22	71.87±0.50⋆
ELECTRA	68.29±1.73	71.14±0.37
Prefix LM	66.86±0.37	69.66±0.78

⋆5-core 100M cluster: all 10 pairwise CIs include zero, within the 0.83pp 9-trial floor. ELECTRA falls within the floor but its CI vs. CLM excludes zero (-0.87 [-1.39, -0.19]). Prefix LM lies outside both criteria (-2.35 [-3.10, -1.76]). Full CIs in Appendix A.5.

token passes (§4.2).

Korean inverts the pattern: the bidirectional−CLM gap (NLU-bidir cluster mean of MLM, Hybrid, AMLM-soft) is -1.33pp at 10M, $+3.34\pm0.33$pp at 30M (every paired pretrain seed positive), and $+1.81\pm0.06$pp at 100M. The 30M midpoint shares its source with KO 10M, controlling for the 100M CC-100 switch. The reversal concentrates on two KLUE subtasks: `klue_nli` (syntax-sensitive; $+15.07\pm0.31$pp at 30M $\rightarrow$ $+2.29$pp at 100M) and `klue_re` (morphology/relation-sensitive; $+1.74$ $\rightarrow$ $+5.37$pp); `klue_sts` and `klue_ynat` go the other way or sit at-noise (Appendix A.18). Causal attribution remains open (see Limitations). At KO 30M, Prefix LM and ELECTRA cluster with CLM below the bidirectional family.

Commonsense scores are near chance with no consistent winner across the four 100M languages and seven objectives (Appendix A.14; per-language tables).

Scaling. The 10M$\rightarrow$100M log-log PPL slope groups objectives into three clusters: *data-hungry* (ELECTRA, Prefix LM), *intermediate* (Span, MLM, AMLM-soft), and *nearly-saturated* (Hybrid, CLM). The P6 120M-parameter probe keeps BLiMP within noise across CLM, Hybrid, AMLM-soft, and Span (Appendix A.11, Table 16).

Metric divergence (exploratory observation). No single objective sits in the top tier of more than three of the five non-chance metrics (Appendix A.12). The mean absolute pairwise Spearman between metrics is $|\rho|$=0.30 over 15 pairs; the BLiMP-to-GLUE pair is $\rho = +0.14$ (Fisher-z CI $[-0.55, +0.81]$, n=7). Phenomenon-level analyses (NPI, ELECTRA, CKA/RSA) appear in Appendices A.22, A.24, A.25, A.10.

6 Discussion

Findings revisited. R1 holds in all four languages and extends EN-only PLL (Salazar et al., 2020) to AMLM-soft and ELECTRA generator PLL. R2 holds in all four 100M languages. R3 is scale-sensitive (§5.3): the MLM-led bidirectional advantage on EN GLUE collapses into the 0.83pp floor by 100M, with NL narrowing, ZH at-noise, and KO 30M departing from the EN trajectory. R4 is a negative scope on commonsense, with cell-to-cell gaps within the binomial-CI floor (Appendix A.14).

Implications for objective choice. Objective choice is conditional on three interacting axes (evaluation paradigm, language, scale); KO 30M is the single observed language $\times$ scale interaction in our four-language sweep, concentrated on syntax- and morphology-sensitive KLUE subtasks (§5.3; boundary-norm probe consistent with a word-boundary account, Appendix A.20). The KO 30M $\rightarrow$ 100M attenuation on `klue_nli` confounds scale-driven narrowing with the BabyBabelLM $\rightarrow$ CC-100 corpus switch (see Limitations). Whether this generalizes to other agglutinative languages (Japanese, Turkish, Finnish, Hungarian) is open; the KO row of Table 19 is a typological hypothesis. Within our sweep, objective choice should be indexed by target and language; English GLUE alone is not a reliable index.

Per-cell observations ($\leq$100M-word, token-matched). Table 19 reports one objective per (target, language) cell. ELECTRA ($\sim$1.5$\times$) and Hybrid ($\sim$1.3$\times$) cost more FLOPs per step than CLM, so a FLOP-budgeted re-ranking is needed within this caveat.

Comparison to prior work. Our scope is $\leq$100M words; Artetxe et al. (2022) preserve the application-conditional bidirectional pattern up to 6.7B in English, and the 100M–1B-parameter gap is unobserved here. Earlier in-regime work (Samuel et al., 2023; Charpentier and Samuel, 2024; Yu et al., 2024; Charpentier et al., 2025) does not provide a matched multilingual objective grid; our EN 100M 5-core within-0.83pp cluster extends this comparison to controlled NL/ZH/KO cells.

7 Conclusion

We presented a matched-budget multilingual sweep of seven pretraining objectives in the data-constrained regime. The English scale trajectory does not predict the Korean midpoint: KO 30M shows a sign-flipped bidirectional advantage on KLUE NLI/RE, which we interpret as a Korean-specific, morphology-sensitive hypothesis consistent with a word-boundary account (Appendix A.20). Objective choice in this regime is conditional on the evaluation paradigm, language, and scale; Table 1 summarizes the per-cell observation. Whether the Korean reversal extends to other agglutinative languages, and how the trajectory continues past 100M parameters, remain open.

Limitations

The sweep is shared-encoder only (30M-param GPT-BERT (Charpentier and Samuel, 2024), with a 120M-param P6 probe); encoder–decoder denoising (T5, BART, UL2) shifts both the parameter-count comparison and the input-output factorization, which would confound the shared-encoder contrast. The Span row uses T5-style sentinels; the SpanBERT boundary-regression formulation (Joshi et al., 2020) is a separate variant, and our ranking applies only to sentinel-based Span. The 100M–1B-parameter regime is outside the scope of this study. ELECTRA results apply only to the shared-encoder setting we test (Appendix A.25).

Per our held-out protocol (§3.2): 10M cells use 3 pretrain seeds at FT 42 (transfer-check stage); 30M/100M cells use 9 trials for the 5-core (endpoint inference) and 3 trials for ELECTRA/Prefix LM. The KO 30M midpoint shares its source with KO 10M and confirms the sign-flip under matched-corpus conditions (Table 20); the KO 100M source switch and ZH 100M epoch setting are detailed in §4.2 and §4.3.

The n=7 Spearman has a wide Fisher-z CI; $\rho = +0.14$ admits metric-specific selection but does not establish a verified null. The 9-trial EN 100M pool has FT-within-pretrain ICC 0.00–0.92 (Appendix A.5); the ELECTRA–CLM contrast is Holm-borderline (Table 10; the shared-encoder bound on ELECTRA is noted above). A fixed 16K byte-level BPE per language serves as the within-language tokenizer control.

The KO boundary-feature probe is robust to boundary definition (Kiwi morpheme-aligned and the BPE whitespace marker both give the same CLM-vs-bidir contrast, positive in every seed at every layer; Appendix A.20). The probe is consistent with a word-boundary account; intervention-based mechanism work is an open direction.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. ExT5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations (ICLR)*.

Stéphane Aroca-Ouellette and Frank Rudzicz. 2020. On losses for modern language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4970–4981.

Mikel Artetxe, Jingfei Du, Naman Goyal, Luke Zettlemoyer, and Veselin Stoyanov. 2022. On the role of bidirectionality in language model pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3973–3985.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: Why not both? In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, pages 1–18.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2023. DUMB: A benchmark for smart evaluation

of Dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32*, pages 13042–13054.

Lukas Edman and Alexander Fraser. 2025. Mask and you shall receive: Optimizing masked language modeling for pretraining BabyLMs. In *Proceedings of the First BabyLM Workshop*, pages 445–453, Suzhou, China. Association for Computational Linguistics.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation. Lm-evaluation-harness, EleutherAI.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, pages 30016–30030.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrama, Francesca Padovani, Francois Meyer, Hai Hu, Julen Etxaniz, Laurent Prevot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2026a. BabyBabelLM: A multilingual benchmark of developmentally plausible training data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3297–3329.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026b. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. PMI-masking: Principled masking of correlated spans. In *International Conference on Learning Representations (ICLR)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. A systematic assessment of language models with linguistic minimal pairs in Chinese. *Preprint*, arXiv:2411.06096. Resource referred to as "ZhoBLiMP" by the authors.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*, pages 1–19.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, pages 50358–50376.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 133–142.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, and 2 others. 2021. KLUE: Korean language understanding evaluation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*. Dataset: https://github.com/KLUE-benchmark/KLUE.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence (Volume 34, Number 5)*, pages 8732–8740.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Sanghoun Song, Kangsan Noh, Eunah Park, Subin Lee, Yebin Lee, and Eunbi Cho. 2025. KBLiMP: How well do language models understand Korean syntactic phenomena? *Language Facts and Perspectives*, 64:265–324. Dataset: https://github.com/kusonglab/kblimp.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22964–22984.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao

Wang, Weijian Xie, and 13 others. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 483–498.

Xinru Yu, Bin Guo, Shiwei Luo, Jie Wang, Tao Ji, and Yuanbin Wu. 2024. AntLM: Bridging causal and masked language models. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 324–331, Miami, FL, USA. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

A Appendix

A.1 Data and Experiment Matrix

Tables 7 and 8 report per-language token counts at 100M and the six-phase experiment matrix (P1–P6). Token budgets differ across languages: EN (Strict-Small / CC-100), NL/KO (BabyBabelLM / CC-100), ZH (BabyBabelLM, sized by a 1.01B-character budget ≈100M ZH-token equivalent). Within each (language, scale) cell, all seven objectives share step count, keeping schedule and token budget matched. The 269 runs group into six phases by role in the argument.

Phase 1 (P1, EN 10M baseline, 21 runs). Seven objectives, three pretrain seeds, with fine-tuning at seed 42. This is where R3 is constructed before any 100M result is consulted.

Phase 2 (P2, cross-lingual 10M held-out, 63 runs). Seven objectives at three seeds on Dutch, Chinese, and Korean. These cells are held out from the rule-setting step and serve as transfer checks for R1 and R2.

Phase 3 (P3, 100M main grid, 84 runs). Four languages times seven objectives times three pretrain seeds, organized into five sub-batches P3.a through P3.e. P3 provides the R1, R2, and R3 endpoints.

Phase 4 (P4, 30M midpoint, 78 runs). P4.a and P4.b train the seven objectives at three seeds on NL/ZH and Korean. P4.c repeats the five core objectives at EN 30M with three pretrain seeds and cross-pretrain fine-tuning. P4.b supplies the KO 30M sign-flip statistic; P4.c anchors the EN GLUE three-point narrowing trajectory.

Phase 5 (P5, EN 10M CC-100 domain check, 7 runs). A robustness check on a different English corpus at matched scale.

Phase 6 (P6, EN 10M 120M-parameter probe, 19 runs). The five core objectives and ELECTRA at three seeds each, plus an ELECTRA seed-789 rerun that replaces the seed-42 generator-PLL outlier under the pre-specified rule.

Table 7: Training data sizes (tokens) and step counts per language at 100M.

Lang	Tokens	Steps	Epochs
EN	150.6M	98,210	10
NL	138.6M	84,620	10
ZH	1.01B	122,876	2
KO	260M	158,680	10

Table 8: Experiment inventory by phase, 269 pretraining runs. 3 pretrain seeds for P1/P2/P3/P4/P6. FT seeds per Table 3.

Phase	Setting	Obj	Runs
P1	EN 10M (BabyLM), 3 seeds	7	21
P2	NL/ZH/KO 10M, 3 seeds	7	63
P3.a	4 langs × 100M, seed 42	7	28
P3.b	4 langs 100M, 5 core, ps 123, 456	5	40
P3.c	EN/NL Prefix LM, ps 123, 456	1	4
P3.d	ZH/KO Prefix LM + EN/KO ELECTRA, ps 123, 456	2	8
P3.e	NL/ZH ELECTRA, ps 123, 456	1	4
P4.a	NL/ZH 30M, 7 obj, 3 seeds	7	42
P4.b	KO 30M, 7 obj, 3 seeds	7	21
P4.c	EN 30M cross-pretrain, 5-core, 3 seeds	5	15
P5	EN 10M (CC-100, domain check), 1 seed	7	7
P6	EN 10M, 120M params, 3 seeds (ELECTRA +1)	6	19
	Total		269

A.2 Hyperparameters and Compute

All runs use AdamW (β_1=0.9, β_2=0.98, ϵ=10^{-6}, weight decay 0.1), peak LR 10^{-3} with 2% linear warmup and cosine decay, bf16, batch 128, sequence 128, 10 epochs (ZH: 2 for the character-count budget). No gradient accumulation. The 30M-param config has 12 Transformer layers, hidden 384, 6 heads, FFN 1536, and a 16K-token BPE vocabulary per language; the 120M (P6) config doubles depth and width.

Objective implementation details. **MLM**: 15% masked tokens, 80/10/10 replace/random/keep

(BERT recipe). **Span Corruption**: contiguous spans (geometric mean 3, range 1–10) at ∼15% density; each span replaced by a sentinel and decoded via the encoder + a generative head. **Prefix LM**: prefix length $k \sim \mathrm{Uniform}(0.25T, 0.75T)$ per sequence; positions $\leq k$ are bidirectional and untraced, positions $> k$ are causal and contribute to the loss (∼50% gradient fraction). **AMLM-soft**: per-token mask probability $p_t \leftarrow \alpha\, p_t + (1-\alpha)\,\mathrm{ema}(\ell_t)$ with $\alpha = 0.99$ (Edman and Fraser, 2025); marginal mask rate held at 15%. **ELECTRA**: 4-layer generator (hidden 192, 6 heads) sharing input embeddings with the discriminator; 15% replaced positions; discriminator-loss multiplier $\lambda = 50$ (Clark et al., 2020). **Hybrid**: two heads on a shared encoder; per step, half the batch uses causal CLM, half uses 15% MLM; losses are summed. **CLM**: standard left-to-right next-token prediction (100% gradient positions).

Total compute is ≈1.0k H100-GPU-hours on a single NVIDIA H100 across P1–P6.

Reproducibility. The anonymized repository README[1] contains the training pipeline, the evaluation harness (including the KBLiMP runner), and the analysis scripts behind the main statistical tables and figures. The lm-evaluation-harness is pinned to v0.4.3 (Gao et al., 2024) (2024-07-01) across all cells so scoring is identical across (language, scale, objective). The camera-ready release will add 246 main-grid HF checkpoints (P1–P4); P5 and P6 are auxiliary controls and will not be released.

A.3 GLUE Fine-Tuning Protocol

EN 7-task evaluation follows Wang et al. (2019b,a): BoolQ (accuracy), MNLI (accuracy), MRPC (F1), MultiRC (accuracy), QQP (F1), RTE (accuracy), WSC (accuracy). The reported **7-task-family average** is the unweighted mean across these 7 tasks.

FT hyperparameters (all conditions): AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}, weight decay 0.01), peak LR 2×10^{-5}, 6% linear warmup with linear decay to 0, batch 32, sequence 128, 5 epochs, bf16. FT seed controls sampling order, classifier-head init, and dropout RNG. **FT seeds**: 9-trial cells use $\{42, 123, 456\}$ (EN 100M all 7 obj; NL/ZH/KO 30M/100M 5-core). Singleton-FT cells (EN 30M; ELECTRA and Prefix LM at NL/ZH/KO) use FT 42. The FT head is linear over [CLS] (sentence) or linear over [CLS] of the concatenated pair (sentence-pair), shared across objectives. Best-epoch dev score reported (no test-set submission).

For ELECTRA, FT uses the discriminator encoder; the generator and binary-classification head are dropped. For Hybrid, FT uses the shared encoder with a single new head.

Cross-lingual GLUE (NL/ZH/KO). Same hyperparameters as EN. Per-task metric is accuracy on all twelve tasks: NL DUMB (de Vries et al., 2023) (sick_nl, dbrd, dutch_cola, sib200_nl); ZH CLUE (Xu et al., 2020) (ocnli, afqmc, tnews, iflytek); KLUE (Park et al., 2021) (re, nli, sts, ynat). For klue_sts (KLUE's primary metric is binary F1 with Pearson's r), dev-set accuracy of the binary similarity head is reported so the 4-task mean is on a common $[0, 1]$ scale. Per-task numbers and splits in Appendix A.18. The 4-task average uses each benchmark's own composition; within-language ranking is the unit of inference.

A.4 Result Provenance

Table 9 maps each main-text claim to its evaluation cell, protocol, and trial count n. Per-cell paths and scoring scripts are in the anonymized `README`. Conventions:

- **Bidir vs. AR (EN trajectory):** AMLM-soft−CLM as the fixed single-objective contrast (comparable across 10M/30M/100M seed protocols); cluster mean (MLM+Hybrid+AMLM-soft)−CLM also reported in Appendix A.18.
- **Bidir cluster (NL/ZH/KO trajectory):** (MLM+Hybrid+AMLM-soft)−CLM, with 30M/100M 5-core cells at 9 trials (3 pretrain × 3 FT); the cluster mean is more stable than any single contrast.
- **Bootstrap:** N=10^4 pairwise resamples of mean differences; CI excludes zero ⇒ significant.
- **Top-tier accounting:** CI-separated for rule-critical multi-seed cells (PPL/GLUE EN 10M); heuristic 2.6pp BLiMP / 1.56pp GLUE noise floor (1.96σ on EN 10M 3-seed pool) elsewhere (Appendix A.12).
- **ELECTRA & Prefix LM at 100M:** 3 pretrain seeds for PPL, BLiMP-family, and cross-lingual GLUE (FT 42 for NL/ZH/KO; 9 trials for EN).

[1] https://anonymous.4open.science/r/objective-choice-data-constrained-B95A/README.md

- **EN 30M (P4.c):** BabyLM Strict subsample; corpus differs from 10M Strict-Small and 100M Strict, so EN 30M PPL is excluded from cross-scale comparison and EN 30M GLUE serves as a trajectory anchor only.

Table 9: Provenance of headline numbers; n counts trials per objective. Per-cell paths in repository `README`.

Claim / cell	Protocol	n
EN 10M PPL	3 pretrain	3
EN 10M BLiMP	3 pretrain	3
EN 10M GLUE	3 pretrain, FT 42	3
EN 100M GLUE (5 core)	3 pretrain $\times$ 3 FT	**9**
EN 100M GLUE (ELECTRA)	3 pretrain $\times$ 3 FT	**9**
EN 100M GLUE (Prefix LM)	3 pretrain $\times$ 3 FT	**9**
NL/ZH/KO 30M, 100M GLUE (5 core)	3 pretrain $\times$ 3 FT	**9**
NL/ZH/KO 30M, 100M GLUE (ELECTRA, Prefix)	3 pretrain, FT 42	3
EN 30M GLUE (P4.c)	3 pretrain, FT 42	3
P6 120M commonsense, 4 obj	zero-shot, 3 pretrain	3
P6 120M commonsense, ELECTRA	zero-shot, single seed	1

Bootstrap 95% CIs use $N{=}10^4$ resamples (Appendix A.5).

A.5 Bootstrap Details

For each pair of objectives, the difference in mean BLiMP or PPL across three English 10M seeds is computed, and the 95% CI is bootstrapped using 10^4 resamples; a difference is significant if its CI excludes zero (BLiMP) or 1.0 (PPL ratios). All BLiMP statistics here use the official lm-evaluation-harness scorer.

Objective	Mean BLiMP (official)	SD
CLM	62.03	0.49
Hybrid	59.20	0.50
Prefix LM	57.83	2.23
MLM	57.59	0.11
AMLM-soft	58.50	2.34
Span	55.77	0.49
ELECTRA	52.47	1.14

Pooled within-objective SD is 1.33pp, giving a $\approx$2.6pp 1.96σ floor (with $n{=}3$, Student-t at $t_{0.05,2}{=}4.30$ gives $\sim$5.7pp). Bootstrap 95% CIs declare 16 of 21 pairwise BLiMP comparisons significant uncorrected; 14 of 21 survive Holm-Bonferroni at $\alpha{=}0.05$ (the demoted MLM–Span and CLM–Hybrid pairs are not rule-critical). The five uncorrected-inconclusive pairs all involve AMLM-soft or Prefix LM, which have the largest seed variance ($\sigma > 2$pp).

GLUE three-seed (English 10M). All seven objectives have GLUE fine-tuning at three pretrain seeds (42/123/456) with a single FT seed (FT 42); $n{=}3$ trials per objective:

Objective	Mean GLUE	SD
AMLM-soft	69.73	0.92
MLM	69.44	0.16
Hybrid	69.28	0.47
Span	69.04	0.22
CLM	68.71	0.43
ELECTRA	68.29	1.73
Prefix LM	66.86	0.37

Pooled within-objective SD on GLUE is 0.80, giving a 1.96σ noise floor of $\approx$1.56pp. Bootstrap 95% CIs ($N{=}10^4$) for top-cluster pairwise differences:

Pair	Δ 95% CI
MLM − AMLM-soft	-0.29 $[-1.31, +0.42]$
MLM − Hybrid	$+0.16$ $[-0.24, +0.63]$
MLM − CLM	$+0.73$ $[+0.25, +1.10]^\dagger$
MLM − Span	$+0.40$ $[+0.15, +0.64]^\dagger$
AMLM-soft − Hybrid	$+0.45$ $[-0.43, +1.49]$
AMLM-soft − CLM	$+1.02$ $[+0.17, +1.93]^\dagger$
AMLM-soft − Span	$+0.69$ $[-0.03, +1.63]$
Hybrid − CLM	$+0.57$ $[+0.01, +1.13]^\dagger$
Hybrid − Span	$+0.24$ $[-0.25, +0.67]$
CLM − Span	-0.33 $[-0.74, +0.12]$

†CI excludes zero.

The bidirectional infill cluster (MLM, AMLM-soft, Hybrid) is above CLM under uncorrected paired-bootstrap intervals (all three contrasts exclude zero). With $n{=}3$ these are quantile intervals rather than large-sample p-values. Under Holm-Bonferroni across 10 pairs, only MLM–CLM and MLM–Span survive; AMLM–CLM and Hybrid–CLM fall to within-noise. The three bidirectional objectives are mutually within-noise. The cluster-level pattern matches the R3 bidir-infill prediction; within-cluster orderings remain inconclusive at three seeds. Bidir−AR contrasts are pre-specified in §3.2.

100M 9-trial bootstrap CI (English). All 7 objectives at EN 100M use 9 trials per objective (3 pretrain $\times$ 3 FT seeds); the 5-core within-noise classification rests on the 10 pairwise differences below, with ELECTRA and Prefix LM contrasts in Table 10. Bootstrap 95% CIs ($N{=}10^4$) for the 5-core 10 pairs:

Pair	Δ 95% CI
CLM − MLM	$+0.09$ $[-0.34, +0.50]$
CLM − Hybrid	$+0.02$ $[-0.38, +0.39]$
CLM − AMLM-soft	-0.09 $[-0.53, +0.32]$
CLM − Span	$+0.13$ $[-0.34, +0.60]$
MLM − Hybrid	-0.07 $[-0.34, +0.17]$
MLM − AMLM-soft	-0.18 $[-0.49, +0.12]$
MLM − Span	$+0.04$ $[-0.34, +0.40]$
Hybrid − AMLM-soft	-0.10 $[-0.36, +0.16]$
Hybrid − Span	$+0.11$ $[-0.23, +0.45]$
AMLM-soft − Span	$+0.22$ $[-0.15, +0.60]$

All 10 pairwise CIs include zero, so the five core objectives at 100M are mutually indistinguishable on GLUE within bootstrap noise; this

matches the 10M→100M narrowing predicted by R3. Mean GLUE over 9 trials per objective: AMLM-soft 72.09 ± 0.35, CLM 72.01 ± 0.59, Hybrid 71.99 ± 0.23, MLM 71.92 ± 0.35, Span 71.87 ± 0.50. Pooled SD is 0.42, giving an empirical 1.96σ floor of 0.83pp; the cluster spread of 0.22pp falls below this floor. Figure 2 plots the 9-trial means with 95% CIs and the full 5×5 pairwise difference matrix.

Holm-Bonferroni-adjusted pairwise contrasts (21 pairs). The 9-trial cluster-bootstrap is extended to all 21 pairs with Holm-Bonferroni at α=0.05 (Table 10). Eleven pairs survive: every contrast involving Prefix LM (6/6) and 5 of 6 involving ELECTRA. *ELECTRA–CLM does not survive Holm* (p_{adj}=0.13), driven by CLM's relatively wide 9-trial SD (0.59pp). All ten within-5-core contrasts fail to reject under Holm, so the "5 core within-noise" classification is robust to multi-comparison correction. ELECTRA is reported as "below each 5-core objective under direct CIs, borderline against CLM under Holm."

Table 10: **EN 100M GLUE pairwise contrasts, Holm-adjusted.** 9-trial cluster-bootstrap (10^4 resamples) at family-wise α=0.05, 21 pairs sorted by $|\Delta|$. ✓ = survives Holm. ELECTRA–CLM is Holm-borderline (p_{adj}=0.13) despite direct CI excluding zero; not counted in the 5-core cluster (Table 6).

Pair	Δ (pp)	95% CI	p_{unc}	p_{Holm}	Holm
AMLM – Prefix	+2.43	[+1.52, +3.11]	0.0000	0.0000	✓
CLM – Prefix	+2.34	[+1.30, +3.25]	0.0000	0.0000	✓
Hybrid – Prefix	+2.33	[+1.42, +2.99]	0.0000	0.0000	✓
MLM – Prefix	+2.25	[+1.33, +2.98]	0.0000	0.0000	✓
Span – Prefix	+2.21	[+1.26, +2.99]	0.0000	0.0000	✓
ELECTRA – Prefix	+1.48	[+0.56, +2.18]	0.0000	0.0000	✓
AMLM – ELECTRA	+0.96	[+0.61, +1.33]	0.0000	0.0000	✓
CLM – ELECTRA	+0.87	[+0.19, +1.44]	0.0116	0.1276	×
Hybrid – ELECTRA	+0.85	[+0.52, +1.19]	0.0000	0.0000	✓
MLM – ELECTRA	+0.78	[+0.35, +1.18]	0.0000	0.0000	✓
Span – ELECTRA	+0.74	[+0.27, +1.20]	0.0022	0.0264	✓
Span – AMLM	−0.22	[−0.66, +0.21]	0.3312	1.0000	×
MLM – AMLM	−0.18	[−0.59, +0.19]	0.3808	1.0000	×
CLM – Span	+0.13	[−0.60, +0.76]	0.6976	1.0000	×
Hybrid – Span	+0.11	[−0.29, +0.54]	0.6042	1.0000	×
Hybrid – AMLM	−0.10	[−0.40, +0.20]	0.4978	1.0000	×
CLM – MLM	+0.09	[−0.60, +0.71]	0.7698	1.0000	×
CLM – AMLM	−0.09	[−0.75, +0.47]	0.8252	1.0000	×
MLM – Hybrid	−0.07	[−0.47, +0.27]	0.7296	1.0000	×
MLM – Span	+0.04	[−0.45, +0.52]	0.8700	1.0000	×
CLM – Hybrid	+0.02	[−0.63, +0.55]	0.9212	1.0000	×

Cluster-bootstrap CIs and ICC (3 pretrain × 3 FT). Per-objective CIs are re-estimated with a cluster-bootstrap that resamples pretrain seeds at the outer level and FT trials at the inner level (N=10^4 resamples, seeded). ICC$(1,1)$ is estimated from a 1-way random-effects decomposition with pretrain seed as the cluster factor:

Obj.	n	Mean (%)	95% CI	ICC	n_{eff}
CLM	9	72.01	[71.40, 72.48]	0.764	3.56
MLM	9	71.92	[71.57, 72.19]	0.481	4.59
Hybrid	9	71.99	[71.79, 72.16]	0.270	5.84
Span	9	71.87	[71.49, 72.23]	0.000	9.00
AMLM-soft	9	72.09	[71.85, 72.32]	0.000	9.00
ELECTRA	9	71.14	[70.84, 71.38]	0.000	9.00
Prefix LM	9	69.66	[69.05, 70.59]	0.916	3.18

ICC ranges from 0.00 (Span/AMLM-soft/ELECTRA) to 0.92 (Prefix LM); CLM is 0.76. n_{eff} ranges from 3.18 to 9.00. The 5-core objectives include zero in their cluster-bootstrap CIs vs. CLM; ELECTRA and Prefix LM exclude zero. The 0.83pp floor at n=9 widens to ≈1.3pp under the conservative n_{eff}=3.18; the classification (6-of-7 within / Prefix LM outside) holds under both bands. Contrasts vs. CLM (pp, 95% CI):

Pair	Δ 95% CI
MLM − CLM	−0.09 [−0.49, +0.34]
Hybrid − CLM	−0.02 [−0.67, +0.72]
Span − CLM	−0.13 [−0.94, +0.69]
AMLM-soft − CLM	+0.09 [−0.49, +0.76]
ELECTRA − CLM	−0.87 [−1.39, −0.19]
Prefix LM − CLM	−2.35 [−3.10, −1.76]

A.6 BLiMP Scorer Sensitivity

Our PLL-based self-computed BLiMP scores differ from the lm-evaluation-harness BLiMP on bidirectional infill objectives:

Objective	Self (PLL)	Official
MLM	51.95	57.47
Span	54.12	55.28
AMLM-soft	44.72	55.82
ELECTRA	52.55	52.78
CLM	71.60	62.36
Hybrid	70.22	59.27
Prefix LM	61.33	60.39

Absolute scores differ across implementations: CLM and Hybrid score ∼9–11 points *higher* on self-PLL, while bidirectional infill (AMLM-soft, MLM, Span) scores *lower*. The ranking roughly agrees within the AR family; cross-family rankings change. Gaps likely reflect implementation differences (temperature handling, span scoring, whitespace tokenization). **All main-text BLiMP results use the official lm-evaluation-harness scorer**; self-PLL values are for transparency only and are not directly comparable across the literature.

A.7 KBLiMP Integration

KBLiMP (Song et al., 2025) is integrated as a custom task in lm-evaluation-harness. KBLiMP covers ten Korean-specific syntactic phenomena with 100 minimal pairs each, for 1,000 minimal pairs total; the release also includes two control sentences per item, which fall outside the minimal-pair protocol

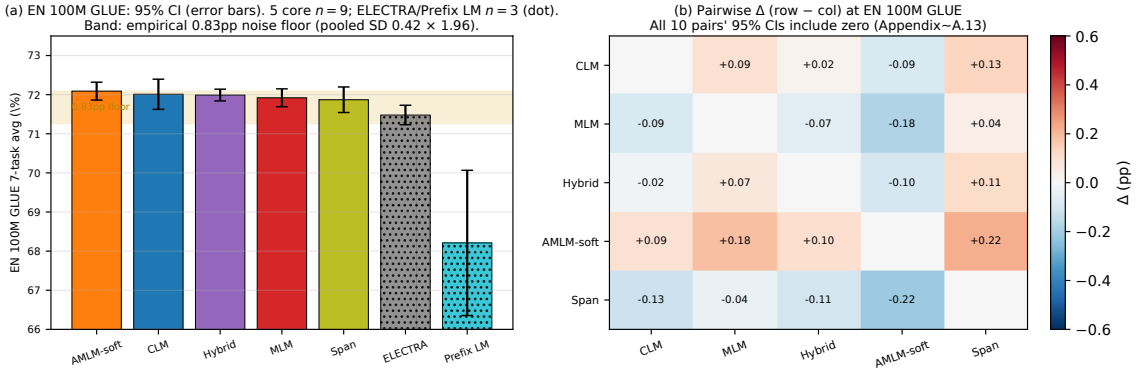

Figure 2: **EN 100M GLUE bootstrap CIs.** (a) 7-obj 95% CI (N=10^4, 9 trials); 0.83pp floor (1.3pp under per-obj n_{eff}). (b) Pairwise Δ for 10 core pairs; all CIs include zero.

and are excluded here. The ten phenomena are supplementary-particle/relative-clause interaction, subject and object case-stacking (*-i/ga*, *-eul/reul*), embedded-clause tense, internally headed relatives (subject and object), complex-NP constraints, comparatives, the *seoldeukhada* alternation, and dative *giving*. Evaluation uses the full 1,000-pair set, and scoring follows the standard BLiMP convention: higher log-likelihood to the grammatical sentence, with accuracy as the fraction correct.

A.8 Per-Language 10M/30M Multi-Seed Results

Three seeds (42/123/456) were additionally trained on NL, ZH, and KO at 10M words to calibrate the noise floor outside English. Tables 11 and 12 report per-language mean±SD; the seven-objective ranking is preserved within noise across all three languages.

Table 11: 10M perplexity, 3-seed mean±SD; ranking preserved across NL/ZH/KO.

Objective	NL	ZH	KO
AMLM-soft	**5.4±1.0**	**7.3±1.6**	**17.9±4.5**
MLM	7.2±0.2	11.6±0.6	25.0±1.6
Span	9.3±0.6	21.6±2.3	36.3±3.5
CLM	18.9±0.5	54.4±1.4	135±5
Hybrid	26.3±0.2	113±3	196±4
Prefix LM	80±4	81±0.5	992±48
ELECTRA	3.1×10^5	2.5×10^4	7.2×10^7

The 10M Dutch BLiMP-NL accuracy (CLM 0.707, Hybrid 0.675) *exceeds* the 100M Dutch result (Table 5, CLM 0.648±0.023) on six of seven objectives (CLM −5.9, Hybrid −11.0, AMLM −8.7, MLM −5.4, Prefix −1.5, ELECTRA −0.9); only Span moves in the other direction (0.611 → 0.617). Both scales use BabyBabelLM Dutch (no CC-100 swap for NL); the predominantly negative drop suggests an extension/diversity effect within BabyBabelLM at 10× scale rather than a corpus-source change. One possible explanation: minimal-pair acceptability may reward developmentally plausible distributional cues that the broader 100M Dutch BabyBabelLM extension may dilute, a hypothesis the present setup does not directly test. The R1/R2 *ranking* is preserved at both scales.

Table 12: 10M BLiMP-family accuracy (EN BLiMP, NL BLiMP-NL, ZH ZhoBLiMP, KO KBLiMP), 3-seed mean±SD. ELECTRA is last in all four (R2).

Objective	EN	NL	ZH	KO
CLM	**0.621**±0.005	**0.707**±0.012	**0.629**±0.027	0.647±0.029
Hybrid	0.592±0.005	0.675±0.009	0.605±0.029	**0.659**±0.025
AMLM-soft	0.597±0.004	0.654±0.013	0.594±0.024	0.634±0.013
MLM	0.576±0.001	0.642±0.018	0.611±0.025	0.613±0.028
Prefix LM	0.578±0.022	0.626±0.036	0.589±0.022	0.571±0.049
Span	0.558±0.005	0.611±0.015	0.620±0.026	0.578±0.039
ELECTRA	0.517±0.015	0.502±0.019	0.527±0.024	0.427±0.036

NL/ZH/KO BLiMP-family at the 30M midpoint (3 seeds). Table 13 reports BLiMP-NL, ZhoBLiMP, and KBLiMP at 30M with 3 pretrain seeds, completing the 10M/30M/100M trajectory for these benchmarks. R2 (CLM tops BLiMP-family) holds at 30M for NL and ZH: NL BLiMP-NL CLM 0.690 ± 0.003 is strict-best, and ZH ZhoBLiMP CLM 0.650 ± 0.008 is strict-best. KO 30M KBLiMP departs from this pattern: AMLM-soft 0.638 ± 0.024 is strict-best, with CLM 0.609 ± 0.015 (−2.9pp) and Hybrid 0.605 ± 0.015 (−3.3pp) below the leader. This 30M KO acceptability ranking is consistent with the same bidirectional-cluster sign-flip seen on KO 30M KLUE FT (§5); both attenuate at 100M, where CLM resumes the BLiMP top (Ta-

ble 5). ELECTRA is last at 30M in all three languages (NL 0.522 ± 0.004, ZH 0.531 ± 0.014, KO 0.428 ± 0.023), consistent with the 100M ranking. The NL 30M $\rightarrow$ 100M drop is again predominantly negative (CLM $0.690 \rightarrow 0.630$, MLM $0.640 \rightarrow 0.568$), reinforcing the BabyBabelLM $10\times$-extension hypothesis discussed above.

Table 13: 30M BLiMP-family accuracy, 3-seed mean±SD. KO 30M KBLiMP reproduces the bidirectional-cluster sign-flip seen on KO 30M KLUE FT.

Objective	NL BLiMP-NL	ZH ZhoBLiMP	KO KBLiMP
CLM	**0.690**±0.003	**0.650**±0.008	0.609±0.015
MLM	0.640±0.021	0.619±0.008	0.530±0.011
Hybrid	0.617±0.015	0.594±0.016	0.605±0.015
Span	0.624±0.013	0.598±0.005	0.514±0.033
AMLM-soft	0.600±0.035	0.608±0.004	**0.638**±0.024
Prefix LM	0.612±0.002	0.575±0.007	0.570±0.023
ELECTRA	0.522±0.004	0.531±0.014	0.428±0.023

A.9 BLiMP Subcategories

Figure 3 shows the 3-seed mean BLiMP accuracy on EN 10M across the 12 BLiMP parent categories (anaphor, argument structure, binding, control-raising, det-noun agreement, ellipsis, filler-gap, irregular forms, island effects, NPI licensing, quantifiers, subject-verb agreement). CLM wins 7 categories; Hybrid wins irregular forms ($+0.0013$ over CLM, within-noise), AMLM-soft wins ellipsis, Span wins filler-gap, Prefix LM wins island effects, and ELECTRA wins NPI licensing. The non-CLM margins are small (Irregular $+0.0013$, Island $+0.0067$, Filler-gap $+0.0192$, Ellipsis $+0.0457$); the largest gap is ELECTRA NPI licensing 59.6% vs AMLM-soft 48.5%, with per-seed SD 5.1pp. These margins are too small for a robust phenomenon-level discrimination claim.

A.10 CKA and RSA

Pairwise CKA (Kornblith et al., 2019) and RSA (Kriegeskorte et al., 2008) are computed between the seven EN-10M checkpoints across seeds 42, 123, and 456, using mean-pooled hidden states at layers 3/6/9/11 over 1,000 BabyLM-dev sentences and reporting the 3-seed mean per (obj, obj) cell. Figure 4 and Table 14 summarize the result.

The bidirectional family (MLM, AMLM-soft, Span) is internally tighter than the AR family at every layer (CKA ≈0.93–0.95 vs. 0.83–0.92; RSA ≈0.87–0.91 vs. 0.53–0.75). Cross-cluster CKA stays 0.88–0.92. ELECTRA is the most distant from the other six and widens with depth: CKA drops $0.862 \pm 0.004 \rightarrow 0.472 \pm 0.004$ from L3 to L9 (3-seed SD), and RSA drops $0.824 \pm 0.007 \rightarrow 0.489 \pm 0.005$, consistent with the category profile in Appendix A.22. The link to downstream behavior is correlational rather than causal.

Table 14: CKA and RSA cluster means by layer (EN 10M, 3-seed mean). *Pure-infill* = MLM/AMLM-soft/Span; *AR/mixed* = CLM/Hybrid/Prefix LM. These representation clusters differ from the NLU-bidir cluster used for cross-lingual GLUE.

Layer	Pure-infill	AR/mixed	Cross	ELECTRA
CKA				
3	0.947	0.919	0.922	0.862
6	0.940	0.897	0.917	0.777
9	0.925	0.903	0.909	**0.472**
11	0.943	0.832	0.875	0.561
RSA				
3	0.913	0.753	0.845	0.823
6	0.910	0.695	0.821	0.769
9	0.874	0.623	0.761	**0.489**
11	0.910	0.533	0.706	0.512

A.11 Scaling Trajectories

Figure 5 visualizes the scaling slopes reported in Section 5.3. Table 15 lists per-objective 3-seed slope, SD, and Student-t 95% CI on the 10M→100M log-log PPL trajectory.

Table 15: EN 10M→100M log-log PPL slope (3 seeds, Student-t 95% CI). Canonical scoring per objective (AR/PLL/generator-PLL). Cluster labels follow threshold cutoffs (see text).

Objective	Slope (mean)	SD	95% CI
Nearly-saturated			
CLM	-0.056	0.015	$[-0.093, -0.019]$
Hybrid	-0.089	0.003	$[-0.097, -0.081]$
Intermediate			
MLM	-0.164	0.012	$[-0.194, -0.133]$
AMLM-soft	-0.173	0.003	$[-0.179, -0.167]$
Span	-0.176	0.023	$[-0.234, -0.119]$
Prefix LM	-0.401	0.084	$[-0.610, -0.192]$
Data-hungry			
ELECTRA	-1.800	0.168	$[-2.217, -1.383]$

Data scaling slopes (EN 10M→100M). ELECTRA's CI $[-2.22, -1.38]$ excludes all others (*data-hungry*). CLM and Hybrid sit near zero (*nearly-saturated*); AMLM-soft, MLM, and Span are *intermediate*. Prefix LM at -0.40 ± 0.08 sits between the intermediate cluster and ELECTRA. The ELECTRA slope is on embedding-tied-generator PLL and is not commensurable with the AR or PLL modes; this slope is used for within-objective scaling only.

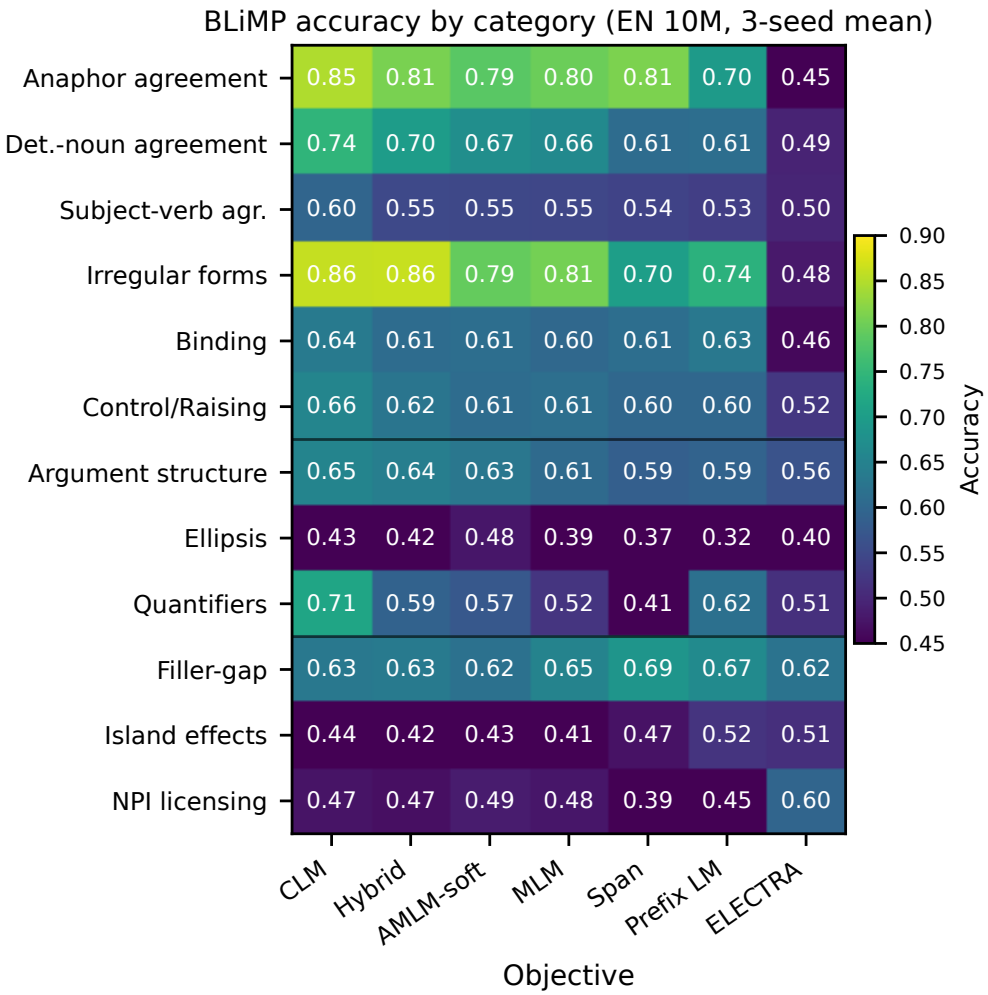

Figure 3: **CLM dominates 7 of 12 BLiMP parent categories at EN 10M.** 3-seed mean over 67 paradigms (CLM wins 7 categories; others 1 each; Gini 0.500). Largest non-CLM gap: ELECTRA NPI licensing $59.6 \pm 5.1\%$ vs AMLM-soft $48.5 \pm 3.0\%$.

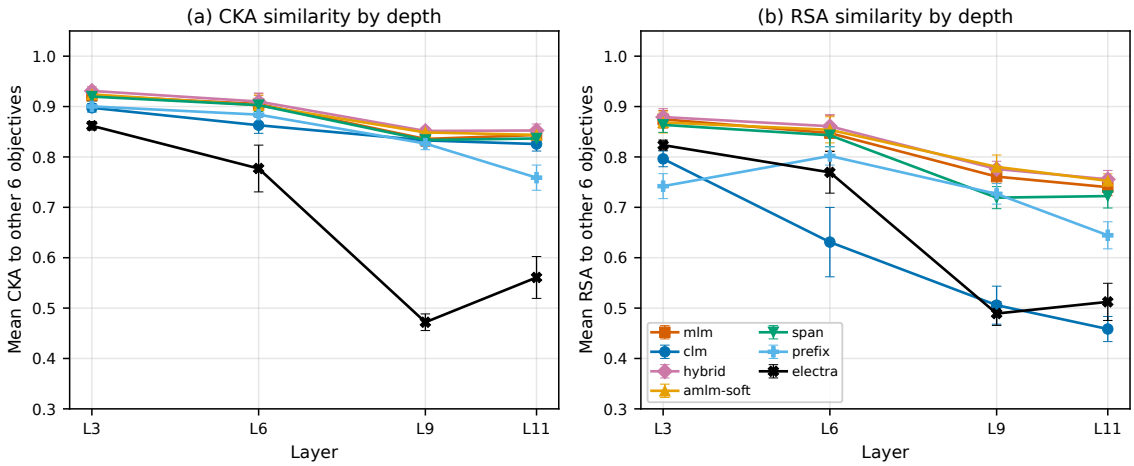

Figure 4: **ELECTRA representations diverge with depth.** Mean pairwise CKA (a) and RSA (b) of each objective to the other six across layers L3, L6, L9, L11; 3-seed mean (error bars = mean across-pair SD per seed). ELECTRA's CKA-to-others drops from 0.862 at L3 to 0.472 at L9; bidirectional infill (MLM/AMLM-soft/Hybrid/Span) clusters above AR (CLM/Prefix LM) at every layer. Full 7×7 matrices per layer in Table 14.

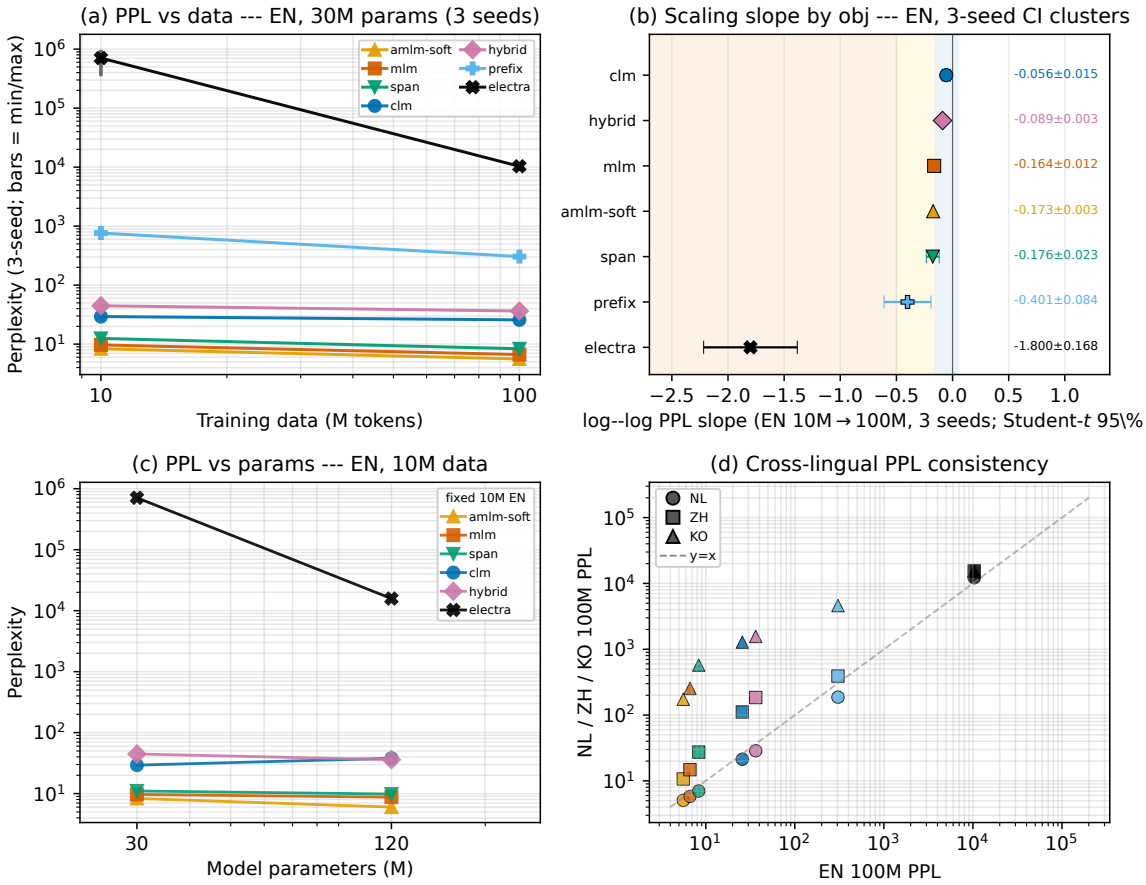

Figure 5: **Scaling clusters by objective (EN).** All panels use 3-seed mean values (seeds 42/123/456; ELECTRA 120M uses 123/456/789 per seed-substitution rule). (a) PPL vs. data (log-log, bars = min/max). (b) Log-log slope with Student-t 95% CI; band colors mark cluster (orange = data-hungry, yellow = intermediate, blue = nearly-saturated): ELECTRA data-hungry; MLM/Span/AMLM-soft intermediate; CLM/Hybrid nearly-saturated; Prefix LM between. (c) P6 PPL vs. params at 10M (Table 16); 6 obj (5-core + ELECTRA). (d) EN vs. NL/ZH/KO 100M PPL (Table 4).

Parameter scaling at fixed 10M. Table 16 reports 3-seed P6 at 30M and 120M parameters; PPL is in each objective's native scoring mode and is not comparable across modes. PLL-scored bidirectional objectives improve from 30M to 120M (AMLM-soft 8.33 → 6.01; MLM 9.69 → 8.74; Span 11.04 → 9.83). AR-scored Hybrid improves (44.65 → 36.00); CLM degrades (29.30 → 37.89), consistent with overfitting under capacity-data imbalance. BLiMP stays within noise across CLM/Hybrid/AMLM-soft/Span. ELECTRA generator PLL improves $\sim 7.1\times10^5 \rightarrow \sim 1.6\times10^4$ (3-seed using 123/456/789 for 120M); the AR/PLL rows are incommensurable with this column.

Table 16: Parameter scaling at fixed 10M EN: 3-seed mean±SD at 30M and 120M. Seeds 42/123/456 except 120M ELECTRA seed-42 (replaced by seed-789 outlier rule; see paragraph below). 30M ELECTRA single-seed. Bold = row winner. †ELECTRA PLL via embedding-tied generator (not comparable to AR/PLL rows).

	PPL		BLiMP	
Objective	30M	120M	30M	120M
AMLM-soft (PLL)	8.33±0.03	**6.01**±0.77	0.585±0.023	0.578±0.004
MLM (PLL)	9.69±0.02	8.74±2.39	0.576±0.001	0.568±0.009
Span (PLL)	11.04±0.20	9.83±0.34	0.558±0.005	0.580±0.004
Hybrid (AR)	44.65±0.08	**36.00**±2.92	0.592±0.005	0.583±0.009
CLM (AR)	**29.30**±0.09	37.89±1.74	**0.620**±0.005	**0.586**±0.010
ELECTRA †	$7.1\cdot10^5$	$\mathbf{1.6\cdot10^4}$	0.528	0.498±0.020

ELECTRA seed-substitution rule (Table 16). The 120M ELECTRA seed-42 generator PLL was 3.9×10^9, more than 100× the median of the other two seeds. ELECTRA's embedding-tied generator PLL has high seed variance across the P6 cells; a uniform >100×-median outlier rule is pre-specified for all P6 cells (the 5-core plus ELECTRA), and only ELECTRA seed-42 triggers it. A fresh seed-789 is retrained, and ELECTRA at 120M is reported from seeds

123/456/789. **The substitution affects only the PPL column; BLiMP, the downstream metric used for R2, is stable across the two seed pools (means 0.498/0.499, SD 0.020/0.014; the pools share seeds 123/456).** For reference, the full ELECTRA 3-seed PLL pool including seed-42 is $7.1 \pm 3.2 \times 10^5$ at 30M parameters and $1.58 \pm 0.27 \times 10^4$ at 120M; neither absolute value nor the outlier's inclusion changes ELECTRA's last-place ranking on any reported metric.

A.12 Top-Tier Objectives Per Metric

The five-metric set is fixed: PPL (PLL mode), PPL (AR mode), BLiMP (EN), MultiBLiMP-family aggregate (NL/ZH/KO), and GLUE (EN 10M rule-critical, R3). Cross-lingual GLUE at 100M is reported separately (Appendix A.18) and not counted. Top tier is (i) CI-separated for rule-critical multi-seed contrasts (PPL/GLUE EN 10M) or (ii) within the heuristic 2.6pp/1.56pp floor elsewhere. CLM achieves the maximum (AR-mode PPL, BLiMP, MultiBLiMP); ELECTRA never enters any top tier.

Robustness to alternative accountings. Counts are accounting-dependent; the qualitative "no universal winner" survives in every alternative tested here. (a) *Add cross-lingual GLUE-equivalent as a sixth slot*: per-objective max stays at 3/6. (b) *Swap EN GLUE 10M for 100M as the GLUE slot*: max becomes 3–4/5 (count moves, the headline survives). (c) *Require CI-separated everywhere*: only rule-critical EN 10M cells remain countable, making the count vacuous. The cross-paradigm $\rho = +0.14$ (Fisher-z CI $[-0.55, +0.81]$) is a weak point estimate under a wide-CI power limit and is consistent with both metric-specific selection and no correlation.

Per-metric top tier and per-objective count. The top objective per metric is AMLM-soft on PLL-mode PPL, CLM on AR-mode PPL, CLM on BLiMP (EN; Prefix LM second at Δ=3.4pp), CLM on MultiBLiMP across all four languages (Span within the floor at ZH and KO), and MLM, AMLM-soft, and Hybrid jointly on EN 10M GLUE. Aggregating EN top-tier appearances over the five non-chance metrics gives CLM 3 (AR-PPL, BLiMP, MultiBLiMP), AMLM-soft 2 (PLL-PPL, GLUE), MLM 1, Hybrid 1, and Span, Prefix LM, and ELECTRA at 0.

A.13 Metric Correlation Matrix

The full Spearman rank correlation between six likelihood-paradigm-comparable metrics is reported over the seven-objective ranking on English 10M. It is computed on three-seed mean scores per objective-by-metric cell, with the analysis referenced in Appendix A.22.

Table 17: Spearman rank correlation across six metrics on EN 10M (n=7). Mean $|\rho|$=0.30; BLiMP↔GLUE ρ=+0.14 (CI $[-0.55, +0.81]$). HSwag/WG/XSC rows are reported for completeness only (R4).

	BLiMP	MBLi.	HSwag	WG	XSC	GLUE
BLiMP	1.00	**0.85**	0.07	−0.11	0.00	0.14
MBLi.		1.00	0.05	−0.07	0.34	0.45
HSwag			1.00	−0.11	0.29	0.00
WG				1.00	**−0.68**	−0.50
XSC					1.00	**0.86**
GLUE						1.00

A.14 R4 Commonsense Numbers

Per-objective accuracy on commonsense benchmarks at 100M (all four languages, 3-seed mean over seeds 42/123/456; chance levels are HellaSwag 0.25, Winogrande 0.50, XStoryCloze 0.50).

Lang	Bench	MLM	CLM	Hyb	AMLM	Span	Pre	ELE
EN	HS	.258	.259	.259	.259	.257	.257	.256
	WG	.496	.515	.498	.508	.506	.509	.494
	XSC	.496	.485	.488	.489	.489	.481	.464
NL	HS	.257	.258	.260	.259	.258	.258	.253
	WG	.500	.496	.502	.509	.504	.505	.506
	XSC	.479	.471	.471	.470	.478	.471	.471
ZH	HS	.255	.256	.257	.254	.255	.256	.255
	WG	.492	.505	.502	.498	.491	.500	.503
	XSC	.461	.466	.461	.462	.464	.462	.461
KO	HS	.256	.253	.255	.258	.255	.256	.254
	WG	.487	.494	.491	.490	.501	.505	.504
	XSC	.470	.474	.475	.478	.473	.469	.477

Worst-to-best within-language gap is at most 2pp on HellaSwag (0.253–0.260, chance 0.25; binomial 95% CI half-width $\approx$ 0.85pp at n=10,042) and Winogrande (0.487–0.515, chance 0.50; CI half-width $\approx$2.75pp at n=1267); on XStoryCloze the within-language gap is at most 3.4pp (0.461–0.496, chance 0.50; CI half-width $\approx$ 2.5pp at n=1511). HellaSwag values are statistically distinguishable from chance per-cell ($z > 5$), but the cell-to-cell gaps remain within 2×CI of each other for all objectives; Winogrande and XStoryCloze cells overlap chance for most objectives. No single objective is the consistent winner across all four languages. R4 is therefore stated as "no consistent cross-language winner" rather than a strict per-cell flat result.

Capacity scaling check (P6, 120M params). To test whether the near-chance commonsense reflects model capacity rather than objective choice, the P6 120M pretrain checkpoints (CLM, MLM, Hybrid, and AMLM-soft at 3 seeds each, plus a single-seed ELECTRA) are re-evaluated on the same three benchmarks. All four core objectives remain near chance: HellaSwag 0.258 ± 0.002 (CLM), 0.257 ± 0.001 (MLM), 0.259 ± 0.002 (Hybrid), 0.258 ± 0.001 (AMLM-soft); Winogrande 0.502 ± 0.006, 0.504 ± 0.004, 0.497 ± 0.006, 0.503 ± 0.004; XStoryCloze 0.480 ± 0.010, 0.484 ± 0.004, 0.488 ± 0.005, 0.482 ± 0.003 (chance 0.25/0.50/0.50). Within-objective deltas relative to EN 30M-parameter values are below 1pp on all three tasks, consistent with R4 across the 30M to 120M parameter range. The modest parameter scaling does not lift commonsense above chance (Figure 6).

A.15 Held-Out Prediction Protocol

Rules R1–R3 are fixed on EN 10M before any 100M result is consulted; they are then evaluated as held-out predictions on (i) NL/ZH/KO and (ii) the 100M scale. R3 is tested on EN at 10M and 100M (3 seeds), and on NL/ZH/KO GLUE-equivalent at 10M and 100M (Appendix A.18). The protocol is not an external preregistration; it serves as an internal held-out test in which rule-critical contrasts and noise-floor thresholds are fixed at the EN 10M stage and reused unchanged on subsequent cells. A rule is falsified if its predicted ranking is reversed beyond the 2.6pp BLiMP / 1.56pp GLUE heuristic floor (6-pair PPL noise for R1). Table 18 summarizes outcomes. No rule-critical contrast was changed after observing 100M results.

Table 18: Held-out predictions vs. 100M outcomes. Detailed numbers and per-cell discussion in §5.

Rule	Prediction	Outcome at 100M	Status
R1	AR tops AR-PPL; bidir tops PLL	Holds in 4/4 languages (both modes)	Supported
R2	CLM tops BLiMP	CLM strict-best in 4/4 (3-seed mean)	Supported
R3	Bidir cluster leads EN GLUE at 10M	EN narrows by 100M; NL attenuates; ZH at-noise; KO inverts at 30M	Heterogeneous
R4	No differentiation on commonsense	Near chance in 4/4; no consistent winner	Supported

A.16 Decision Guide Table

Table 19 reports one objective per (target, language) cell. Where evidence within a cell is within-noise, the most-supported single objective in this test set is picked, with within-noise alternatives noted in a footnote. The four columns instantiate Germanic fusional (EN, NL), Sinitic analytic (ZH), and Koreanic agglutinative (KO) typologies.

Table 19: Decision table: one objective per (evaluation target, language). Picks are token-matched (not FLOP-matched); scope is $\leq$100M words. Floors and FLOP ratios are summarized in §3 and Appendix A.2.

Target \ Lang.	EN	NL	ZH	KO
BLiMP, 100M	CLM	CLM[a]	CLM[b]	CLM[c]
PPL (AR), 100M	CLM	CLM	CLM	CLM
PPL (PLL), 100M	AMLM	AMLM	AMLM	AMLM
GLUE, EN 10M	MLM[d]	—	—	—
GLUE, 100M	5-core[e]	CLM[f]	CLM[g]	bidir[h]
Commonsense, 30–120M	—	—	—	—

[a,b,c] Span within floor at ZH/KO; Span and Prefix LM outside at NL.
[d] MLM Holm-significant; AMLM/Hybrid Holm-borderline.
[e] 5-core within the 0.83pp floor.
[f,g] CLM strict-best; bidirectional cluster within the floor.
[h] NLU-bidir−CLM $+3.34/+1.81$pp at 30M/100M (Appendix A.20).
"—" not applicable or all objectives near chance.

A.17 EN Domain Robustness

EN data is BabyLM Strict-Small at 10M and BabyLM Strict at 100M. As a domain specificity check, all seven objectives were re-trained on EN CC-100 10M (P5, seed 42). CC-100 GLUE 7-task averages: AMLM-soft 71.43, MLM 71.26, ELECTRA 70.43, Hybrid 70.34, Span 70.33, CLM 68.73, Prefix LM 68.28; the bidirectional cluster leads and Prefix LM is last, mirroring the EN 10M BabyLM ranking. Per-objective BabyLM/CC-100 deltas are within the 1.56pp GLUE floor for the 5-core, so the EN 10M cluster ranking is domain-stable. This check does not control for the KO 100M corpus switch (BabyBabelLM $\rightarrow$ CC-100); morphology and corpus shift remain entangled at KO 100M. The EN narrowing itself is within BabyLM (Strict-Small to Strict) and is not a domain switch.

A.18 Cross-lingual GLUE (10M→30M→100M)

Four sequence-classification tasks per language are evaluated at 10M, 30M, and 100M (NL DUMB (de Vries et al., 2023), ZH CLUE (Xu et al., 2020), and KO KLUE (Park et al., 2021), all compatible with single-head fine-tuning). The 10M cells use 3 pretrain seeds at FT 42 for the 5 core obj. NL/ZH/KO 30M and 100M cells use 3 pretrain $\times$ 3 FT (9 trials per cell) for the 5 core obj, and 3 pretrain $\times$ FT 42 (3 trials per cell) for ELECTRA and Prefix LM.

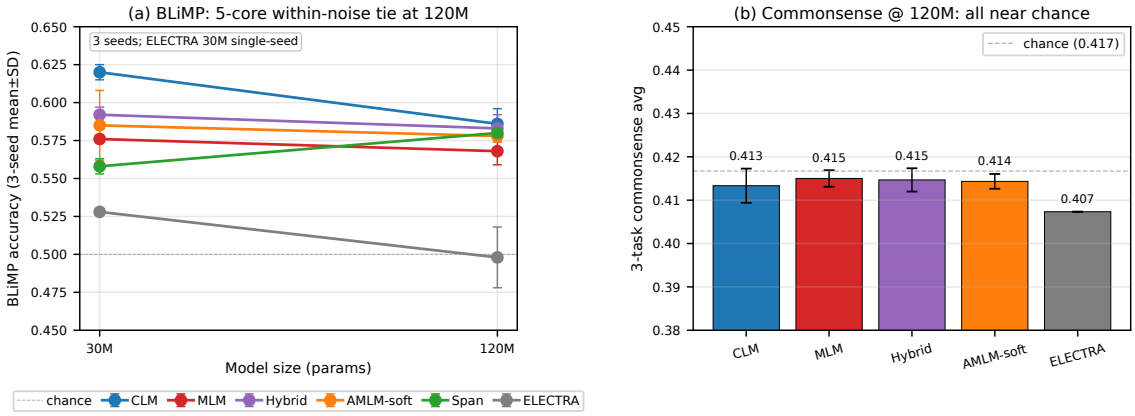

Figure 6: $4\times$ **params at fixed 10M: BLiMP tightens at 120M; commonsense does not separate.** P6 scaling (EN 10M, 3 seeds). (a) BLiMP: CLM/Hybrid/AMLM-soft/Span within noise at 120M; ELECTRA below cluster. (b) 3-task commonsense mean (chance 0.417); all near chance.

EN 30M midpoint (multi-seed). The 5-core objectives were trained at EN 30M with 3 pretrain seeds and FT 42 (15 trials), anchoring the EN 10M→100M GLUE narrowing with a 3-point trajectory. At 30M, the 5 core means cluster within ∼2.5pp (Hybrid top, Span last; full per-objective numbers in Table 20). The representative AMLM-soft−CLM gap evolves $+1.02 \rightarrow +0.89 \rightarrow +0.08$pp across 10M, 30M, and 100M; the cluster mean (MLM+Hybrid+AMLM-soft)−CLM is $+0.77 \rightarrow +0.97 \rightarrow -0.02$pp. The 30M midpoint sits within the 1.56pp floor of both endpoints. The EN 30M training corpus is the BabyLM Strict 30M subsample (CHILDES-heavy, distinct from Strict-Small at 10M and Strict at 100M), so EN 30M PPL is excluded from cross-scale comparison. The EN CC-100 10M domain check (Appendix A.17) shows the 5-core GLUE cluster ranking survives one 10M domain shift within the 1.56pp floor; whether this generalizes to the BabyLM Strict 30M CHILDES shift is not directly tested. The EN 30M GLUE midpoint is therefore read as a directional anchor under this caveat.

Three distinct trajectories appear; the NLU-bidir cluster is the MLM/Hybrid/AMLM-soft mean (Span and Prefix LM are reported separately). **NL**: CLM leads at all three scales, with NLU-bidir−CLM at $-3.41 \rightarrow -1.24\pm1.13 \rightarrow -0.62\pm0.65$pp (10M → 30M → 100M); the gap attenuates across the three points and 4 bidir obj enter the 100M top tier within noise. **ZH**: all 7 obj within ∼1.5pp at every scale (already converged at 10M); the 30M gap of -0.64 ± 0.12pp is tight and slightly negative. **KO**: the 10M ranking is within-noise but inverts at 30M. Per-pretrain-seed NLU-bidir−CLM gaps are $+3.00$, $+3.66$, and $+3.36$pp (all positive), mean $+3.34\pm0.33$pp (paired-t 95% CI $[+2.52, +4.16]$). The effect persists at 100M ($+1.81\pm0.06$pp). KO 30M shares its source with KO 10M (BabyBabelLM), reducing the chance that the 100M sign-flip is solely a CC-100 artifact; the effect concentrates on syntax/morphology-sensitive subtasks. At KO 30M, Prefix LM (60.38 ± 0.65) and ELECTRA (62.78 ± 0.82, FT 42 only) cluster with CLM, consistent with the AR side of the bidir−AR contrast extending to partially-unidirectional objectives.

Per-task NLU-bidir−CLM (9 trials per cell): at 100M, klue_re (relation, morphology-heavy) $+5.37 \pm 0.39$pp, klue_nli (syntactic-semantic) $+2.29\pm0.31$pp, klue_sts (lexical) -0.27 ± 0.93pp, klue_ynat (topic) -0.17 ± 0.10pp. At 30M: klue_nli $+15.07\pm0.31$pp, klue_re $+1.74\pm0.64$pp, klue_ynat -0.55 ± 0.36pp, klue_sts -2.89 ± 0.48pp. The NLU-bidir advantage concentrates on syntax- and morphology-sensitive tasks at both scales and vanishes on lexical tasks; the klue_nli gap is largest at 30M and attenuates by 100M, consistent with R3-style narrowing within these subtasks. The boundary activation-norm probe (Appendix A.20; 3 seeds, 500 sentences, Kiwi and Ġ) is consistent with a word-boundary account; this aligns with prior tokenization-side evidence that morpheme/eojeol-aware preprocessing helps Korean NLU (Park et al., 2020), though the mechanisms (representation-level boundary norms vs. input-side morpheme segmentation) are distinct: CLM has the lowest boundary/non-boundary norm ratio (∼0.53 Kiwi, ∼0.43 Ġ); the NLU-bidir cluster sits at ∼0.93/∼0.78; the paired gap is positive

Table 20: Cross-lingual 4-task mean per language (NL DUMB, ZH CLUE, KO KLUE; per-task metric in Appendix A.3). 10M: FT 42 (3 pretrain seeds). 30M/100M: 9 trials for 5-core (3 pretrain × 3 FT); 3 trials for ELECTRA/Prefix LM. Mean±SD. **Bold** = per-cell top; ∗ = within the per-cell empirical floor (1.56pp for n=3 10M cells, 0.83pp for n=9 30M/100M 5-core cells). EN AMLM−CLM trajectory: +1.02 → +0.89 → +0.08pp.

	NL			ZH			KO		
Objective	10M	30M	100M	10M	30M	100M	10M	30M	100M
MLM	70.20±1.72	78.86±0.89∗	79.90±0.81∗	61.95±0.36∗	61.94±0.20∗	63.28±0.27∗	62.74±0.55∗	65.06±0.50∗	**75.55**±0.53
CLM	**73.36**±0.40	**80.26**±0.70	**80.67**±0.39	**62.50**±0.16	**62.57**±0.21	**63.31**±0.25	**63.78**±1.02	61.56±0.46	73.61±0.43
Hybrid	71.13±1.46	79.21±1.06∗	80.49±0.29∗	62.09±0.18∗	61.91±0.22∗	63.23±0.13∗	62.31±0.97∗	**65.37**±0.45	75.45±0.29∗
AMLM-soft	68.53±0.87	78.99±0.54∗	79.76±0.55∗	62.35±0.18∗	61.96±0.31∗	63.14±0.13∗	62.31±0.58∗	64.28±0.73∗	75.23±0.44∗
Span	68.12±1.13	79.31±0.85∗	79.48±0.37∗	61.46±0.25∗	61.42±0.30∗	62.96±0.29∗	61.74±0.68	63.02±1.12	75.26±0.49∗
Prefix LM	—	78.88±0.73∗	76.03±0.51†	—	62.37±0.14∗	63.21±0.08†∗	—	60.38±0.65	72.90±0.40
ELECTRA	—	72.92±1.08	74.11±1.03	—	58.94±0.51	60.32±0.17	—	62.78±0.82	69.53±0.64
NLU-bidir − CLM	−3.41	$-1.24^{\pm 1.13}$	$-0.62^{\pm 0.65}$	−0.37	$-0.64^{\pm 0.12}$	$-0.09^{\pm 0.26}$	−1.33	$+3.34^{\pm 0.33}$	$+1.81^{\pm 0.06}$

in every seed at every layer under both measures. The EN BabyLM/CC-100 check (Appendix A.17) suggests but does not isolate KO corpus shift at 100M (Appendix A.19).

A.19 Korean Corpus Robustness

KO 100M uses CC-100 while KO 10M uses BabyBabelLM (Jumelet et al., 2026a). To check whether a corpus-quality confound affects R1's "ranking preserved" claim, the seven-objective PPL ranking is compared at the two KO scales:

Objective	10M PPL	100M PPL	Rank
AMLM-soft	17.9	173.7	1 / 1
MLM	25.0	253.8	2 / 2
Span	36.3	575.4	3 / 3
CLM	135.0	1,281	4 / 4
Hybrid	196.3	1,564.9	5 / 5
Prefix LM	992.1	4,612	6 / 6
ELECTRA	–	15,452	7 / 7

(KO 10M = BabyBabelLM; KO 100M = CC-100. 100M cells: 3-seed mean (seeds 42/123/456). 10M Prefix LM/ELECTRA cells use the seed-42 PPL value from the 10M evaluation; ELECTRA's PLL is incomparable with the AR-mode column at 10M (omitted with "–"). Ranking preserved across the simultaneous corpus and scale change.)

The ranking is preserved exactly across the simultaneous corpus and scale change. Absolute PPLs are higher at 100M because the held-out validation set is larger and more diverse (CC-100 web text); R1's ranking claim holds despite the simultaneous corpus and scale shift; the two contributions are not separated by this 2-point comparison.

A.20 KO Boundary-Feature Probe

The probe asks whether the KO bidirectional advantage at 100M is consistent with a word-boundary account, without any intervention. For seven objectives at 100M (three pretrain seeds), per-position L2 hidden-state norms are measured on 500 BabyBabelLM-dev sentences under two boundary definitions: **Kiwi morpheme-aligned** (primary, via the Kiwi Korean morphological analyzer[2]; 29% of non-special BPE tokens) and **Ġ eojeol-aligned** (sanity check, using the BPE word-initial marker; 55% of tokens). The two measures are complementary (Kiwi∩Ġ/Kiwi = 35%).

Table 21: KO 100M boundary/non-boundary L2 norm ratio per layer; 3-seed mean±SD (500 dev sentences). **Kiwi** (morpheme, primary) and **Ġ** (BPE marker, sanity). NLU-bidir−CLM paired gap positive in every seed at every layer under both definitions.

Objective	Layer 3	Layer 6	Layer 9	Layer 11
Kiwi morpheme-aligned (primary)				
CLM	0.53 ± 0.02	0.53 ± 0.03	0.53 ± 0.02	0.53 ± 0.02
MLM	0.90 ± 0.00	0.89 ± 0.03	0.88 ± 0.02	0.88 ± 0.03
Hybrid	0.93 ± 0.03	0.92 ± 0.03	0.91 ± 0.04	0.91 ± 0.04
Span	0.94 ± 0.02	0.93 ± 0.01	0.93 ± 0.03	0.96 ± 0.04
AMLM-soft	0.95 ± 0.06	0.94 ± 0.05	0.93 ± 0.05	0.93 ± 0.05
ELECTRA	0.98 ± 0.03	0.96 ± 0.05	0.95 ± 0.05	0.97 ± 0.01
Prefix LM	0.91 ± 0.02	0.92 ± 0.02	0.96 ± 0.01	0.99 ± 0.01
Ġ eojeol-aligned (sanity)				
CLM	0.43 ± 0.02	0.42 ± 0.03	0.43 ± 0.02	0.44 ± 0.02
MLM	0.77 ± 0.02	0.75 ± 0.06	0.74 ± 0.04	0.74 ± 0.04
Hybrid	0.77 ± 0.04	0.78 ± 0.04	0.74 ± 0.05	0.74 ± 0.05
Span	0.81 ± 0.01	0.79 ± 0.04	0.80 ± 0.04	0.86 ± 0.03
AMLM-soft	0.83 ± 0.06	0.82 ± 0.07	0.81 ± 0.07	0.82 ± 0.06
ELECTRA	0.98 ± 0.04	0.95 ± 0.07	0.94 ± 0.09	0.95 ± 0.09
Prefix LM	0.83 ± 0.01	0.82 ± 0.01	0.87 ± 0.01	0.95 ± 0.01

CLM has the lowest ratio at every layer (Kiwi ∼0.53, Ġ ∼0.43); norms at boundary tokens are 47–57% smaller than at boundary-internal tokens. The bidirectional infill cluster (MLM, Hybrid, AMLM-soft) sits at ∼0.88–0.95 (Kiwi) / ∼0.74–0.83 (Ġ); Span is similar. The paired NLU-bidir−CLM gap is +0.40 → +0.37 across L3→L11 under Kiwi (and +0.37 → +0.34 under Ġ), positive in every seed at every layer under both measures. ELECTRA's ratio is near 1.0 under both; Prefix LM sits between the cluster and ELECTRA. The pat-

[2] Kiwi (kiwipiepy) v0.17, https://github.com/bab2min/kiwipiepy.

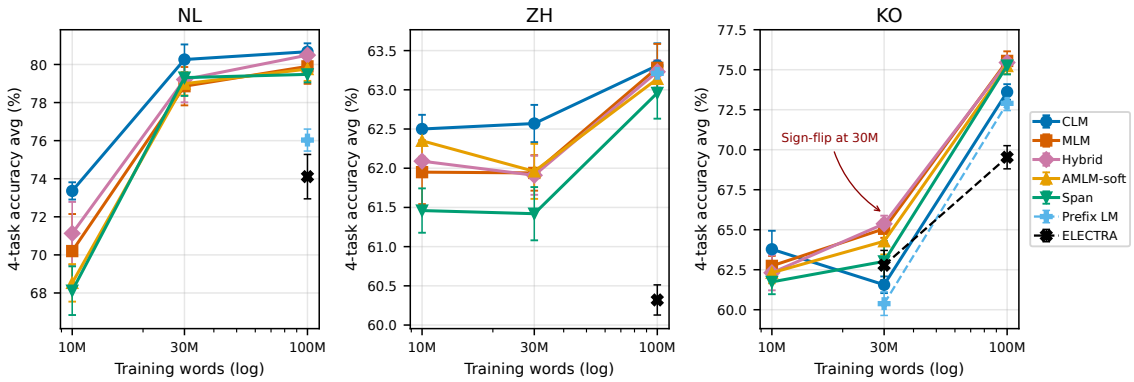

Figure 7: **NL narrows, ZH at-noise, KO inverts at 30M.** Cross-lingual 4-task scores at 10M/30M/100M (3 pretrain seeds); KO 30M shares its source with KO 10M.

tern aligns with the KLUE 100M per-task gradient: the largest bidir advantages are on klue_re and klue_nli (where boundary information matters), while the advantage vanishes on lexical tasks (klue_sts, klue_ynat). The probe is non-interventional (hidden states at 4 sampled layers) and is robust to boundary definition.

A.21 Tokenizer Fertility

Why is Korean PPL ten-fold higher? KO 100M PPLs (AMLM 173, CLM 1,265) are an order of magnitude above EN/NL/ZH. Tokenizer fertility (Table 22) shows KO has half the chars/token of EN (1.93 vs. 3.38) at fixed 16K vocab, fragmenting content into roughly 2× tokens. The absolute PPL is incomparable; the *ranking* is preserved (Appendix A.19).

Does tokenizer fertility confound BLiMP accuracy? KO mean sentence length at the BPE token level (7.2 tokens, Table 22) is half of EN's (15.1), so KO minimal-pair scoring averages over fewer per-position likelihoods. Within-language objective ranking is robust to this (the 7 objectives share the same tokenizer per language), but cross-language BLiMP-family absolute values are not directly comparable: a ranking-of-objectives interpretation is appropriate while an absolute-difficulty comparison is not. The BLiMP-family benchmarks across languages also differ in paradigm coverage, so token-count effects are entangled with benchmark design.

A.22 NPI Licensing Probe

ELECTRA is last on the BLiMP average yet leads NPI licensing across 3 seeds (59.6±5.1%; AMLM-soft 48.5±3.0%, near chance); Island effects has

Table 22: Tokenizer fertility on held-out validation (16K byte-level BPE per language, trained on the 10M corpus). For KO, "word" = orthographic eojeol.

Lang	chars/token	tokens/word	avg sent (tokens)
EN	3.38	1.59	15.1
NL	3.72	1.60	19.2
ZH	1.49*	–*	long doc.
KO	**1.93**	1.75	7.2

*Chinese has no whitespace word separation; chars/token here is per-Han-character, and "words" are not well-defined.

ELECTRA (51.0±4.6) within seed variance of Prefix LM (51.7±2.8). Across the 12-category aggregation (3-seed mean; Figure 3): CLM 7, Hybrid/AMLM-soft/Span/Prefix LM/ELECTRA 1 each (winner-Gini 0.500). Non-CLM wins have within-category gaps of 0.001–0.046 vs seed variance, except NPI. Reported as a phenomenon-level observation without Bonferroni adjustment.

A.23 EN GLUE Narrowing Mechanism

The EN 10M to 100M GLUE delta is +3.30pp for CLM and +2.36 to +2.85pp for the remaining six objectives. CLM gains slightly more than the rest; ELECTRA and Prefix LM gain alongside the bidirectional cluster but start from lower 10M baselines. Spearman across 7 objectives gives info-flow $\rho = +0.42$ and gradient-position fraction $\rho = +0.73$. EN 10M three-seed bootstrap with the official scorer gives 16 of 21 BLiMP pairs and 6 of 6 within-mode PPL pairs excluding zero (Appendix A.5), with pooled SD of 1.33pp.

A.24 ELECTRA Scoring Robustness

R1 scores ELECTRA via PLL using its generator, which shares input embeddings with the discriminator but not full encoder weights (Clark et al., 2020).

As a robustness check, ELECTRA's discriminator confidence is also reported (mean $P(\text{original})$ on clean text, where 50% is chance and 100% is perfect):

Setting	Disc. confidence
EN 10M (mean of seeds 123/456)	90.4%
EN CC-100 10M	88.1%
NL 10M	91.5%
ZH 10M	91.8%
KO 10M	84.2%

Discriminator confidence is 84–92% across languages, well above chance. This number is an internal clean-text diagnostic, not directly comparable to the GLUE or generator-accuracy metrics reported by Clark et al. (2020). ELECTRA's R1 last-place reflects scoring-paradigm mismatch (likelihood vs. discrimination); the model itself converges normally.

A.25 Discriminator-Only Control

To isolate ELECTRA's discriminator objective from the learned-generator design, a discriminator-only variant is trained on EN 10M with the same configuration as full ELECTRA but with the learned generator replaced by uniform-random vocab sampling. The setting samples from non-special tokens at each masked position, with three pretrain seeds (42/123/456), 15,875 steps, batch 128, and an identical optimizer/schedule. The discriminator still performs replaced-token detection but is no longer trained against an adversary that produces semantically plausible replacements.

Setting	BLiMP mean±SD	n
Full ELECTRA (EN 10M)	0.517±0.015	3
Discriminator-only (EN 10M)	0.513±0.001	3
Δ (full − disc-only)	$+0.004$ (within noise)	—
CLM (EN 10M, reference)	0.620±0.003	3

Removing the learned generator changes BLiMP by less than the within-objective seed SD, with σ_{full}=0.015 and σ_{disc}=0.001, and the full versus disc-only difference of 0.4pp is over an order of magnitude below the roughly 10pp gap to CLM. This rules out the strongest version of the generator-design-carries-the-signal interpretation, since ELECTRA's BLiMP last-place persists under random replacement and is therefore consistent with the PLL and discriminator scoring-paradigm mismatch (Section 3, Appendix A.24) rather than a representational limitation that the generator would supply. Disc-only is not a complete control for the shared-encoder confound, since the GPT-BERT encoder is the same in both variants; a native discriminator-only architecture with a separate encoder and no shared embeddings remains future work.