ETH zürich

Contrastive Decoding for Synthetic Data Generation in Low-Resource LM

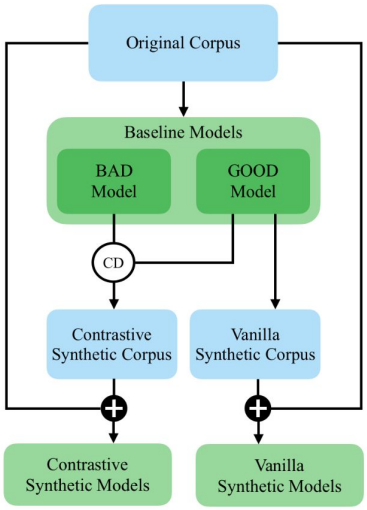
Jannek Ulm[1], Kevin Du[1], Vésteinn Snæbjarnarson[1,2]
[1]ETH Zurich; [2]University of Copenhagen

1 Motivation

LLMs face **high-quality data scarcity**; BabyLM asks what's possible with **100M words**. Synthetic text is a candidate fix—but risks **hallucination, bias, collapse** if done naively. Question: can **contrastive decoding (CD)** turn synthetic text into a **useful training signal** under a strict budget?

2 Method Overview

Train baselines: ~100M-param LLaMA-style models on **TinyBabyLM (≈100M words)**. **Pick GOOD** (best checkpoint); define **BAD** (weaker variant). **Generate synthetic corpora (≈100M tokens)** via: **Vanilla** ancestral sampling (NO-CONTRAST); **Contrastive Decoding (CD)**: promote GOOD's preferences, subtract BAD's. **Train from scratch** on **70/30 real/synthetic**, identical hyperparams to baselines. **Evaluate** on the BabyLM suite + PPL; report **mean–max** per task with paired bootstrap.



3 Key Results

All synthetic regimes > baseline; **CD-Early-500** is best **overall**: **+4.90% μΔREL** across tasks. With **Top-k=200**, CD reaches **+5.69% μΔREL** at ~unchanged PPL. **Vanilla** has the **lowest PPL** (**23.56**), but **CD** wins on **reasoning-heavy** tasks: **BLiMP Supplement** (65.10), **Entity Tracking** (30.38), **EWoK** (53.80), **WUG** (70.55), **Eye Tracking** (4.42). **Head-to-head (CD vs. Vanilla)**: CD gains **+7.3%** (Entity), **+8.2%** (WUG), **+1.2%** (EWoK); Vanilla slightly better on **PPL** and **BLiMP**.

Name	$\mu_{\Delta\text{REL}}$ ↑	Perplexity↓	BLiMP↑	BLiMP Supp.↑
Baseline	-	24.46±0.10	71.03±0.27	64.10±0.60
No-Contrast	2.96%	**23.56±0.11***	**72.09±0.17***	64.83±0.73
No-Contrast-V-Head	0.66%	24.33±0.10*	71.67±0.24*	64.86±0.74
CD-Early-500	**4.90%**	23.73±0.10*	71.72±0.19*	**65.10±0.60***

Name	Entity Tracking↑	EWoK↑	WUG↑	Reading↑	Eye Tracking↑
Baseline	27.82±1.18	53.18±0.28	66.90±2.47	1.76±0.22	3.85±0.31
No-Contrast	28.14±1.75	53.17±0.30	64.67±1.66*	**1.91±0.25**	4.31±0.33*
No-Contrast-V-Head	25.47±1.40*	53.03±0.31	66.67±1.58	1.76±0.23	4.32±0.33*
CD-Early-500	**30.38±0.65***	**53.80±0.29***	**70.55±2.32***	1.79±0.22	**4.42±0.32***

4 Why CD Helps (Intuition)

It's the **contrastive scoring**, not just a GOOD-head mask .**NO-CONTRAST–VHEAD** helps PPL a bit but **hurts Entity Tracking** (–8.45%); CD's gains come from the **subtraction against BAD**.

5 BAD Model (Amateur) Ablation

Tried three BADs: **earlier checkpoint**, **smaller model**, **dropout proxy**. **Earlier checkpoint** (e.g., step 500) is **strongest** and simplest operationally. Best overall config: **CD-Early-500 + Top-k=200 → +5.69% μΔREL**.

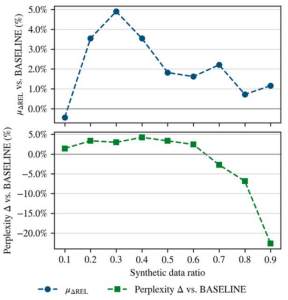
Name	$\mu_{\Delta\text{REL}}$ ↑	Perplexity↓
BASELINE	-	24.46±0.10
NO-CONTRAST	2.96%	**23.56±0.11*** (3.68%)
NO-CONTRAST-Top-k-200	3.65%	23.65±0.10* (3.29%)
CD-Small-20	3.55%	23.73±0.14* (2.96%)
CD-Drop-0.7	3.29%	24.06±0.13* (1.65%)
CD-Early-500	4.90%	23.73±0.10* (2.98%)
CD-Early-500-Top-k-200	**5.69%**	23.77±0.10* (2.80%)

6 Decoding Truncation Ablation

Top-k (light) helps most; **k=200** = best aggregate (**+5.69%**). **k=100** yields **strongest Entity Tracking** (**+19.02%**) and best **EWoK (+1.44%)**. **Top-p** less reliable; can reduce μΔREL despite small PPL wins (for Vanilla).

5 Synthetic Data Mixing Ratio Ablation

30% synthetic in batches → **best overall μΔREL (+4.90%)**. **40% synthetic** → **lowest PPL (≈23.42)** in our sweep.



5 Conclusion

Use **CD-generated corpora** when your target tasks need **reasoning, tracking, or world knowledge**. Use **vanilla sampling** when you want **lower perplexity / grammatical regularities**. In our sweep, the simplest and most effective amateur was an **earlier checkpoint** of the same model; **30% synthetic** data ratio was best overall, **40%** minimized perplexity.