Enhancing Linguistic Competence of Language Models through Pre-training with Language Learning Tasks

ACL 2026 (Oral) - Harbor H-I, Session 4, Oral Session C: Linguistic theories, Cognitive Modeling and Psycholinguistics 1, Sun. July 5 16:00-17:30.

Atsuki Yamaguchi

Collaborators:
Maggie Mi Nikolaos Aletras

29 May 2026
@Cambridge

University of Sheffield

Self-introduction

Atsuki Yamaguchi

Sheffield NLP Group

A final-year PhD candidate (Thesis submission due in March 2027)

Research Interests

Cross-lingual transfer and efficient language modelling

- NLP Researcher at Hitachi, Ltd. (2021-2023)
- ACL Rolling Review (ARR) Support Team member (2024-)
- EACL 2026 Student Research Workshop (SRW) Chair

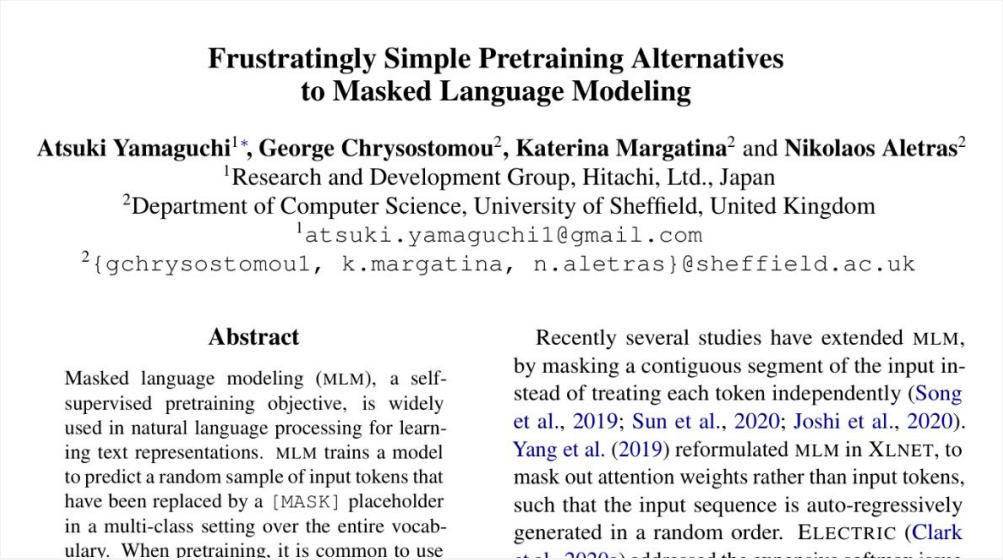
💼 I’m looking for faculty/research scientist positions!

Talk Structure

1. How project started
2. The motivation for L2T
3. The L2T Framework
4. Results & Analysis
5. Discussion & Future Directions

Exploring Efficient Encoder Pre-training

Back then…



Frustratingly Simple Pretraining Alternatives to Masked Language Modeling

Atsuki Yamaguchi[1]*, **George Chrysostomou**[2], **Katerina Margatina**[2] and **Nikolaos Aletras**[2]
[1]Research and Development Group, Hitachi, Ltd., Japan
[2]Department of Computer Science, University of Sheffield, United Kingdom
[1]atsuki.yamaguchi1@gmail.com
[2]{gchrysostomou1, k.margatina, n.aletras}@sheffield.ac.uk

Abstract

Masked language modeling (MLM), a self-supervised pretraining objective, is widely used in natural language processing for learning text representations. MLM trains a model to predict a random sample of input tokens that have been replaced by a [MASK] placeholder in a multi-class setting over the entire vocabulary. When pretraining, it is common to use

Recently several studies have extended MLM, by masking a contiguous segment of the input instead of treating each token independently (Song et al., 2019; Sun et al., 2020; Joshi et al., 2020). Yang et al. (2019) reformulated MLM in XLNET, to mask out attention weights rather than input tokens, such that the input sequence is auto-regressively generated in a random order. ELECTRIC (Clark

Frustratingly Simple Pretraining Alternatives to Masked Language Modeling

Yamaguchi et al., EMNLP 2021

- Re-evaluating the dominance of MLM
- Introducing efficient pre-training tasks

Exploring Efficient Encoder Pre-training

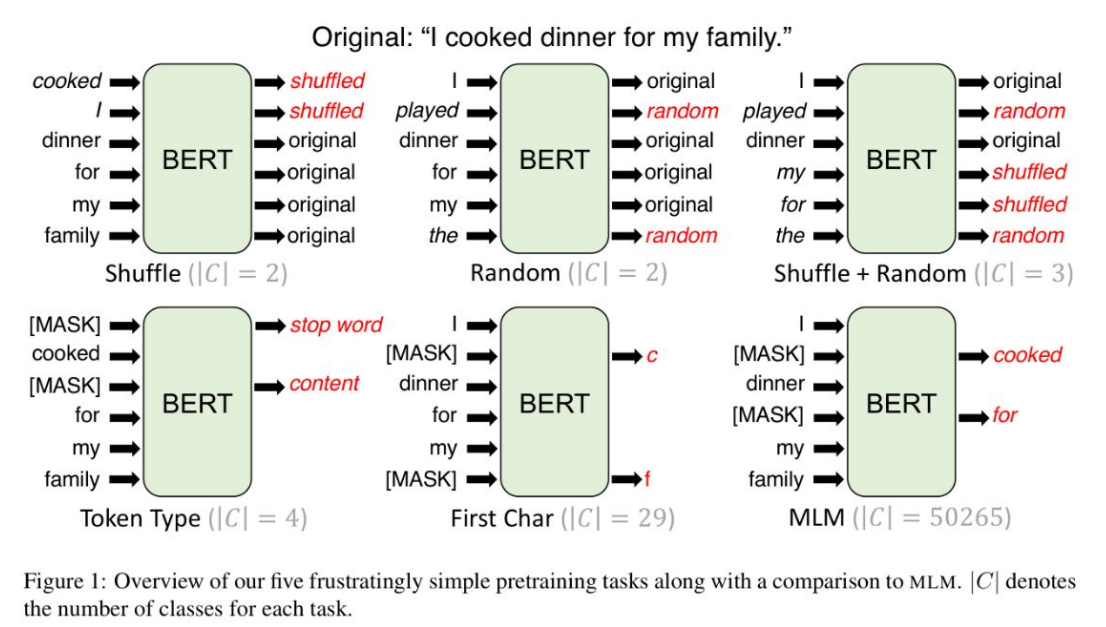


Figure 1: Overview of our five frustratingly simple pretraining tasks along with a comparison to MLM. $|C|$ denotes the number of classes for each task.

Key Observation

MLM wastes ~85% of tokens.

Strategy

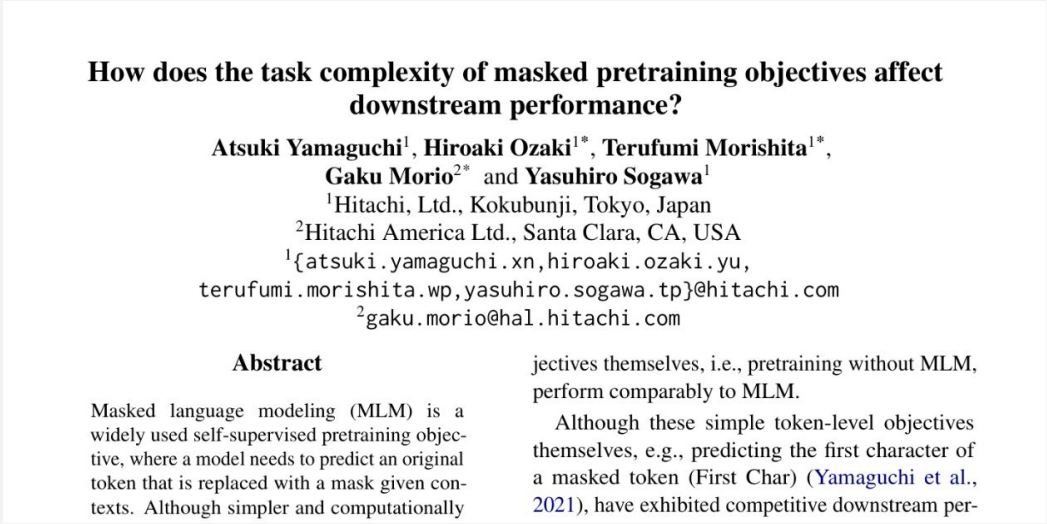
Focus on simple dense token-level tasks.

Takeaway

You don't necessarily need bidirectional masking to induce rich lexical representations; simple auxiliary token classifications can suffice.

Understanding Efficient Encoder Pre-training

Following up my EMNLP ‘21 paper: Why certain token-level tasks outperform others?



How does the task complexity of masked pretraining objectives affect downstream performance?

Atsuki Yamaguchi[1], **Hiroaki Ozaki**[1*], **Terufumi Morishita**[1*], **Gaku Morio**[2*] and **Yasuhiro Sogawa**[1]
[1]Hitachi, Ltd., Kokubunji, Tokyo, Japan
[2]Hitachi America Ltd., Santa Clara, CA, USA
[1]{atsuki.yamaguchi.xn,hiroaki.ozaki.yu,
terufumi.morishita.wp,yasuhiro.sogawa.tp}@hitachi.com
[2]gaku.morio@hal.hitachi.com

Abstract

Masked language modeling (MLM) is a widely used self-supervised pretraining objective, where a model needs to predict an original token that is replaced with a mask given contexts. Although simpler and computationally jectives themselves, i.e., pretraining without MLM, perform comparably to MLM.

Although these simple token-level objectives themselves, e.g., predicting the first character of a masked token (First Char) (Yamaguchi et al., 2021), have exhibited competitive downstream per-

(Yamaguchi et al., ACL Findings 2023)

Key Finding

Downstream performance correlates with auxiliary task complexity (target space cardinality).

Takeaway

Pre-training must use high-entropy distributions to force the model to learn better representations

Then, decoder models came!

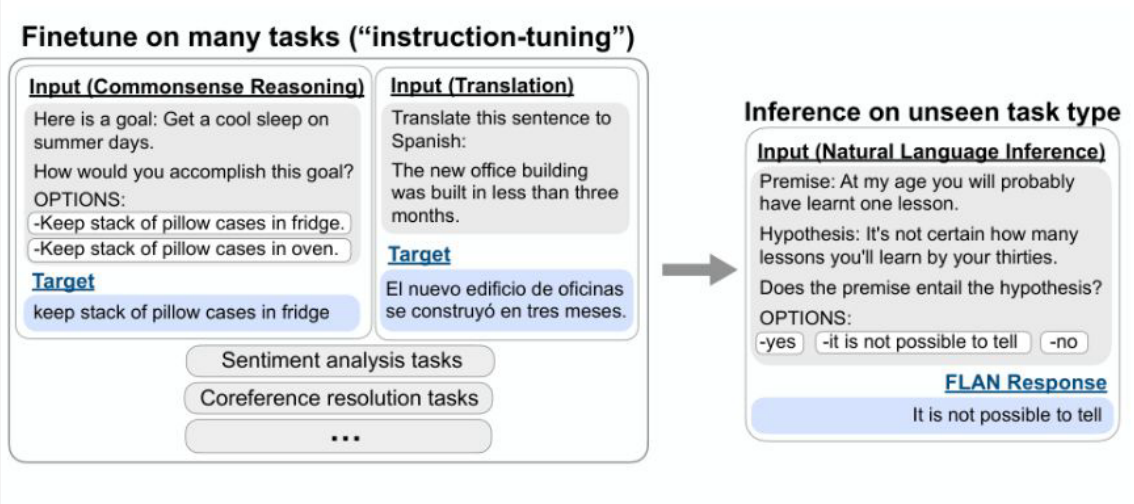
Rapid Change:
The landscape shifted quickly as encoder pre-training studies became obsolete.

Why the Shift? → Flexibility

Native *zero-shot* and *few-shot* support removes the need for task-specific heads.

Rise of "Data"

Data-centric studies (like Flan →) gained attention.



Finetuned language models are zero-shot learners (Wei et al., ICLR 2022)

One day in January 2025

The Proposal

Can we adapt our EMNLP ‘21 encoder tasks for decoder models to boost performance?

Nikos

Example: Task-Specific Conversion

Input: "Cat sat on the mat"

Task: Shuffle
→ **New Input:** "sat Cat mat on the"
→ **Label:** "Cat sat on the mat."

This forces LMs to reason over the full context?

Beyond rote memorisation

Standard causal LM on raw text:

➕ Facilitates the learning of world knowledge and reasoning

➖ But, it does not optimise for linguistic competence.

Bender et al. (2021): LMs act as **"Stochastic Parrots"**

- Just matching surface-level string frequencies rather than parsing underlying grammar.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 (Bender et al., FAccT 2021)

How to explicitly optimise linguistic competence?

Our approach: Mimicking human language acquisition!

Simulates how humans learn*: explicit error correction, morphological awareness training, and discourse completion.

These are called Language Learning Tasks (L2T).

*Afra Alishahi. 2011. Computational Modeling of Human Language Acquisition, first edition. Synthesis Lectures on Human Language Technologies. Springer Cham.
*Jennifer Culbertson and David Adger. 2014. Language learners privilege structured meaning over surface frequency. Proceedings of the National Academy of Sciences, 111(16):5842–5847.

How does L2T work?

Standard CLM

Phase 1: Input Processing

Raw documents (unstructured) data

↓

Phase 2: Training

Train on raw data with CLM

Formulation:

```
Input: ([Task Instruction])
    + [Corrupted/Partial Text]
Target: [Original Text]
    or [Answer]
```

L2T Framework

Phase 1: Input Processing

Raw documents (unstructured) data

↓

Phase 2: L2T Transformation

Generate prompt-completion pairs (character-, word-, sentence-, and discourse-level tasks) - 14 tasks

↓

Phase 3: Training

Train on **L2T** + raw data with CLM

Character-level Task

Enhance morphological awareness and discourage surface-level matching.

Task 01: Character Counting (Char Count)

INPUT `Calculate char count: "The cat."`

TARGET **7**

Task 02: Masked Character Recovery (Masked Char)

INPUT `A single g+++ld aiguill++++++te is worn ...`

TARGET `A single gold aiguillette is worn ...`

Task 03: Space Restoration (Space)

INPUT `Ilikecats`

TARGET `I like cats`

Task 04: Typo Correction (Typo)

INPUT `indiaiduals with a first-degree relatiye`

TARGET `individuals with a first-degree relative`

Word-level Task

Promote structural inference over sequential statistics.

Task 05: Word Last (Last)

INPUT [Text] Options: A. concluding phrase B. decoy
TARGET A. concluding phrase

Task 06: Masked Word

INPUT I [MASK] a student.
TARGET I am a student.

Task 07: Random Correct (Random)

INPUT Sea am hungry.
TARGET I am hungry.

Task 08: Shuffle Restore (Shuffle)

INPUT white loops buckles and permitted ...
TARGET white loops and buckles permitted...

Task 09: Token Type Count (Token Type)

INPUT Count Digit in: "3 dogs and 2 cats."
TARGET 2

Sentence-level Tasks

Promote structural inference over sequential statistics.

Task 10: Sentence Deletion (Deletion)

INPUT Hakama is usually made of gray silk. (PDF) STOP, THINK, SPOT FAKE NEWS. Women wear them only for specific occasions ...

TARGET (PDF) STOP, THINK, SPOT FAKE NEWS

Task 11: Sentence Reordering (Reorder)

INPUT [S3] Demands for change may come. [S1] CTIN 534 introduction ... [S2] This course will introduce students ...

TARGET [S1] → [S2] → [S3]

Discourse-level Tasks

For global coherence and ambiguity resolution

Task 12: Fill in the middle (Fill Middle)

INPUT Prefix: "The case was identified ... " Suffix: " ... Brazil has seen deaths." Complete middle:

TARGET "although researchers said there are indications ... "

Task 13: One-Word Prefix Generation (One)

INPUT Generate text starting with: "These"

TARGET These routes allow visitors to locate…

Task 14: Complete the latter half (Half)

INPUT …Compassionate nurses who are experienced in helping children through

TARGET these tests will start an intravenous line and help administer medications to relax…

Experimental Setup

Evaluating L2T on 500M and 1B architectures with a controlled 100B token budget in two scenarios:

Disjoint Scenario

- Simulates standard scaling where unique source text is plentiful.
- Documents sampled from mutually exclusive subsets
- **Tests general scaling efficiency.**

Shared Scenario

- Simulates data-constrained settings using the exact same base texts for raw and L2T data.
- Decouples task structure from data diversity.
- **Tests if L2T improves linguistic inductive bias without new data.**

Evaluation Framework

Baselines Pre-train models on raw data

- Disjoint Raw: 100B unique tokens.
- Shared Raw: 42B unique tokens, trained multiple times to reach 100B quota for fair comparison.

Linguistic Competence Evaluation (BLiMP)

- Measures core linguistic knowledge by measuring log-likelihood of minimal pairs.
- Domains: Semantics, Morphology, Syntax.
- Example: Irregular form ✅ *Aaron broke the bicycle.* ❌ *Aaron broken the bicycle.*

General Benchmarks Confirm if L2T tasks complement CLM.

- Reading Comprehension (RC): RACE, SciQ, LogiQA
- Commonsense Reasoning (CR): ARC-Easy, COPA, OpenBookQA, SIQA, HellaSwag
- LAMBADA (Language Modeling), ReCoRD (RC + CR)

Results: L2T boosts linguistic competence

Findings

L2T: better overall gains across settings.

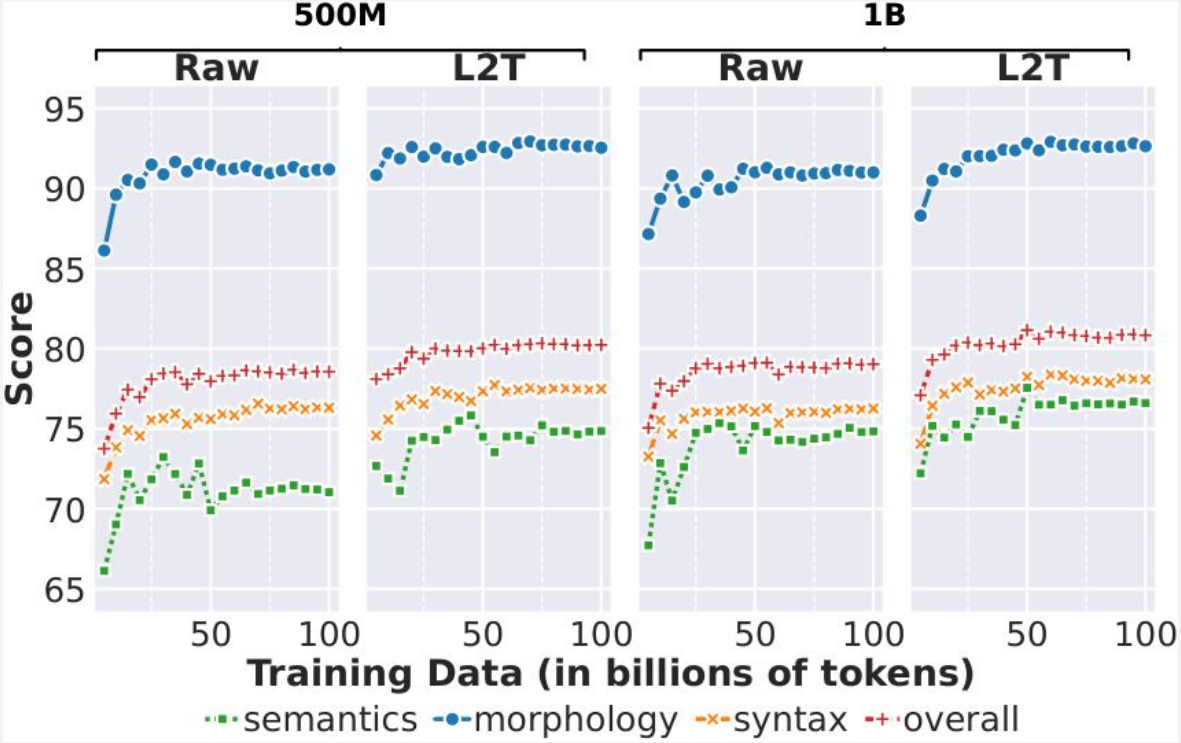
Substantial gains (up to +11.3 points in Island Effects)

No improvement for saturated phenomena (determiner-noun agreement, ellipsis)

	Data		Morphology	Syntax		Overall
			DN Agr	Ellips	Island	
500M	Disjoint	Raw	93.1	87.1	63.0	78.6
		L2T	92.4	86.0	70.8	80.2
	Shared	Raw	93.6	88.5	60.6	78.1
		L2T	93.4	86.2	68.1	80.9
1B	Disjoint	Raw	94.5	86.7	60.2	79.0
		L2T	93.3	86.5	71.5	80.8
	Shared	Raw	94.6	86.5	61.7	78.9
		L2T	92.7	84.5	68.6	81.2

Table: Linguistic competence on BLiMP. **Green highlights** indicate performance of L2T models over Raw baselines.

L2T accelerates linguistic competence acquisition



L2T: Acquire linguistic competence at a highly accelerated rate within the first **20%** of token exposure.

→ L2T boosts learning in "*the window of maximal development*" (Shah et al., EMNLP 2024)

Development of Cognitive Intelligence in Pre-trained Language Models (Shah et al., EMNLP 2024)

How about downstream performance?

Disjoint Configuration

General reasoning performance remains stable with access to unique documents.
(-0.9 to 0.0 on average)

Shared Configuration

Impact of L2T varies by model scale.

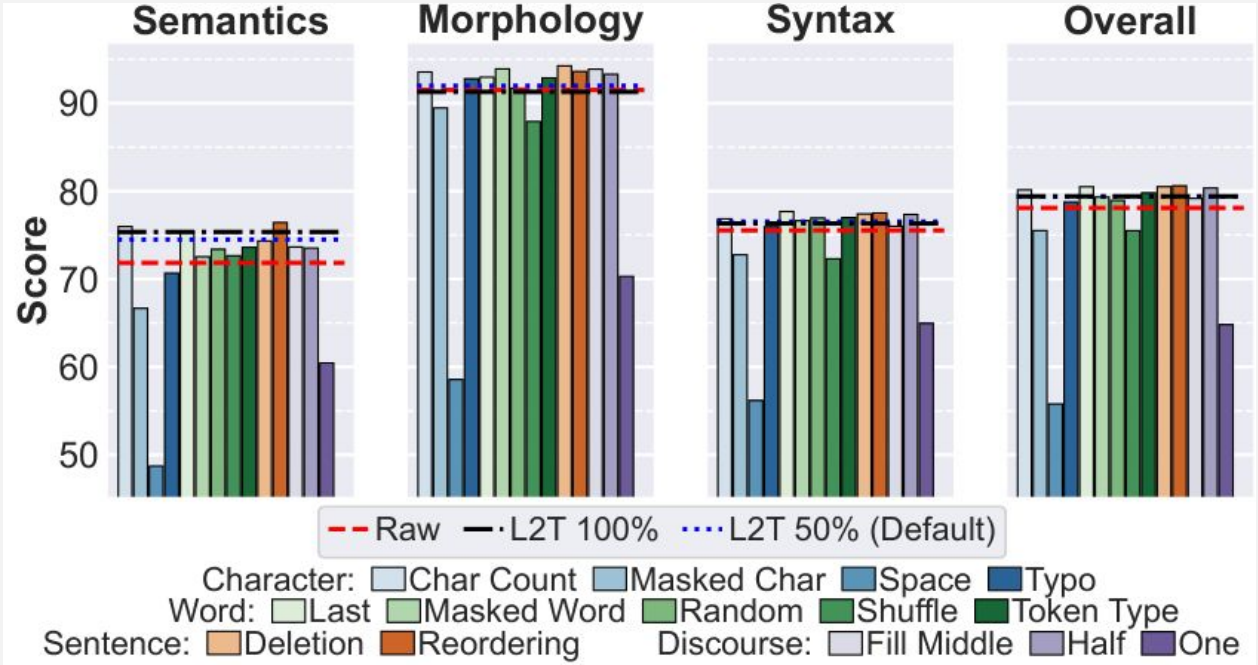
- **500M Model:** Remains the same
- **1B Model:** Drops -1.4 points (up to ARC: -4.2)
- ***Shows tension between linguistic structure learning (L2T) and factual consolidation in larger models.***

	Data		CR ARC	CR Overall
500M	Disjoint	Raw	57.4	46.4
		L2T	56.4	45.5
	Shared	Raw	56.6	45.7
		L2T	57.7	45.7
1B	Disjoint	Raw	60.4	47.3
		L2T	58.4	47.3
	Shared	Raw	60.6	47.8
		L2T	56.4	46.4
Random guessing			25.0	27.5

Table:General benchmark performance. Green denotes better performance of L2T (ours) over Raw.

Which L2T task is important?

Combined L2T framework elicits complementary strengths across linguistic phenomena.



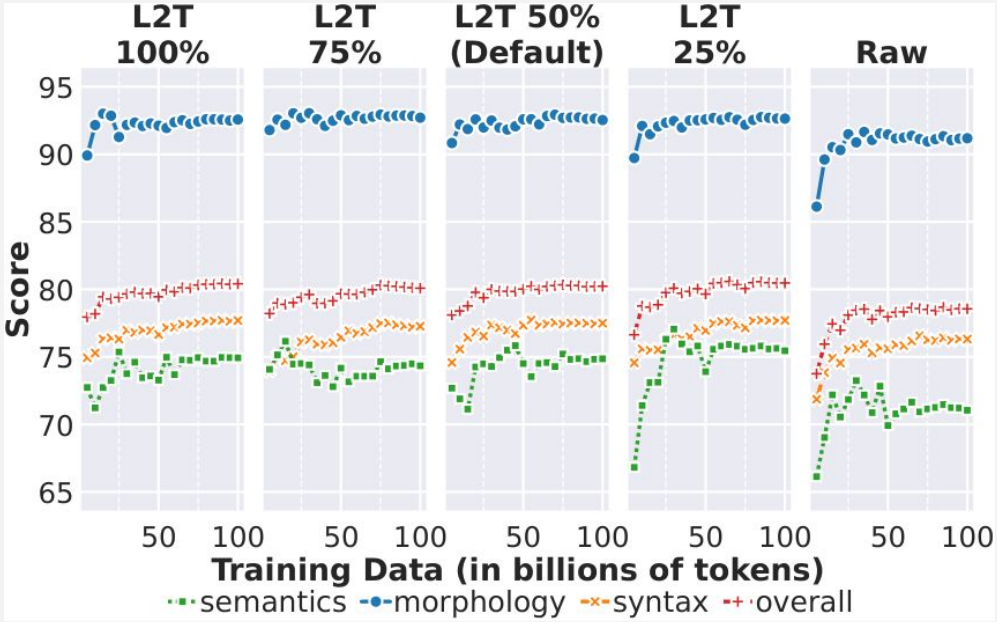
High Impact:
Nine tasks (e.g., *Char Count*, *Reordering*) consistently outperform Raw baseline

Low Impact:
Character-level tasks (e.g., *Space*, *Masked Char*) underperform, likely due to unstable standalone signals.

Mixing ratio highlights tension in general knowledge

Linguistic Competence

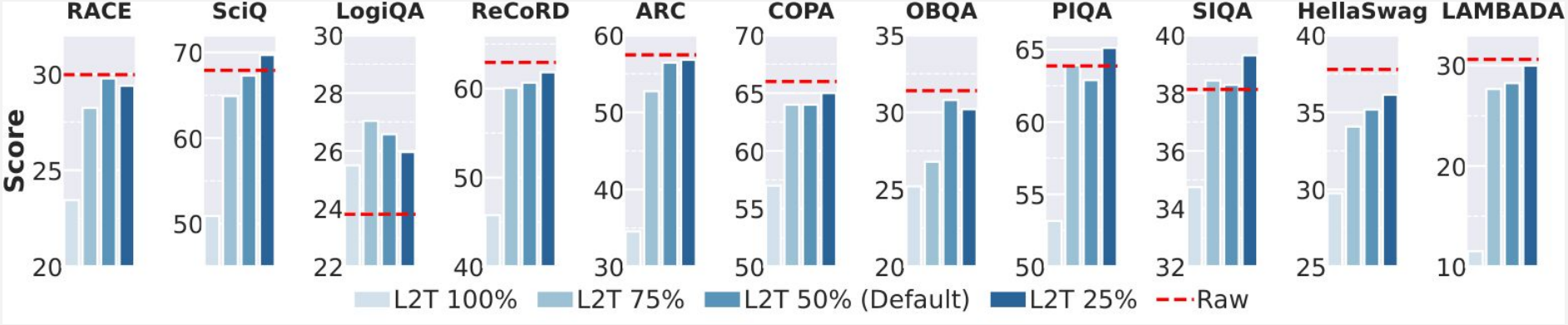
- Minor final differences (max delta 1.1 in semantics)
- Higher L2T ratios accelerate learning (5B tokens)



Mixing ratio highlights tension in general knowledge

General Benchmarks

- Raw text is essential for broad knowledge and reasoning!
- A strong correlation (0.67 to 1.0) between raw text percentage and performance.
- Adding even 25% raw text (L2T 75%) mitigates gap



Discussion: Performance trade-off

➕ L2T provides superior linguistic competence.

➖ A subtle drop in factual memorization.

- In Shared setting, replacing raw sequence repetitions with L2T data reduces commonsense reasoning performance.

💡 Our synthetic data pre-training study* shows a similar trend, *lowering performance on generalised commonsense reasoning tasks*.

*How Can We Synthesize High-Quality Pretraining Data? A Systematic Study of Prompt Design, Generator Model, and Source Data, (Niklaus et al., 2026)

Discussion: How can we effectively use L2T?

L2T accelerates linguistic competence early but has diminishing returns after the "window of maximal development." A curriculum format could be beneficial:

- **Stage 1:** Heavily interleave structured L2T tasks early to build strong grammar.
- **Stage 2:** Transition to pure CLM sequences to maximise factual ingestion, knowledge consolidation, and world-reasoning capacity.

This is future work!

Discussion: Is L2T effective for non-English languages?

- L2T is English-only.
- Performance impact of L2T tasks may differ substantially in other languages.
  - For instance, scriptio continua languages (e.g., Thai, Japanese) which lack word spaces.
- Testing is now possible using FineWeb2 (Penedo et al., COLM 2025) and MultiBLiMP (Jumelet et al., TACL 2026).

This is also future work!

FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language (Penedo et al., COLM 2025)
MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs (Jumelet et al., TACL 2026)

University of Sheffield

Thank you!