- MVA 2025/2026 Deep Learning Project -
- Making Transformers Solve Compositional Tasks (Ontanon et al., 2022) -

Conducted by: Younes BOUKACEM & Yassine AIT-SAID

I. Context

Natural language processing (and AI more broadly) has been transformed by large language models (LLMs), which achieve SOTA performance including Out-Of-Distribution (OOD) generalization.

At the heart of LLMs is the Transformer architecture and its attention mechanism → **Ontanon et al. (Google Research, 2022)** propose to study the OOD generalization capacity of small Transformers trained on toy synthetic tasks:

Reverse:
Input: 1 3 3 7 2 [END]
Output: 2 7 3 3 1 [END]

Duplication:
Input: 1 3 5 7 2 [END]
Output: 1 3 5 7 2 1 3 5 7 2 [END]

Addition:
Input: # # # 3 6 7 [SEP] # # 1 4 9 1 [END]
Output: # # 1 8 5 8 [END]

Cartesian:
Input: 1 2 3 [SEP] a b [END]
Output: 1 a [SEP] 2 a [SEP] 3 a [SEP] 1 b [SEP] 2 b [SEP] 3 b [END]

Figure 2 - Some synthetic tasks from Ontanon et al. (Excerpt)

Specifically, they analyze the impact of different architectural choices on the Transformer OOD generalization.

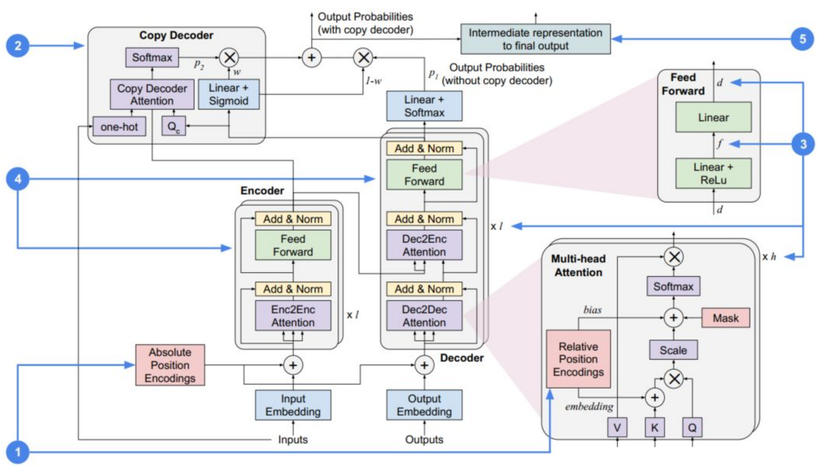


Figure 3 - Transformer components studied by Ontanon et al. (Excerpt)

Limitations:

1. Only studies Encoder-Decoder Transformers, which are actively superseded by Decoder-Only Transformers in LLMs.
2. The OOD generalization tests were conducted by extending the inputs size in the synthetic tasks, but did not consider extending the inputs number.
3. Learnable absolute positional encoding was not considered in the study (only sinusoidal), while the relative positional encodings were learnable embeddings.

II. Defining Our Synthetic Tasks

We adapt two tasks from Ontanon et al. to our multi-inputs setup:

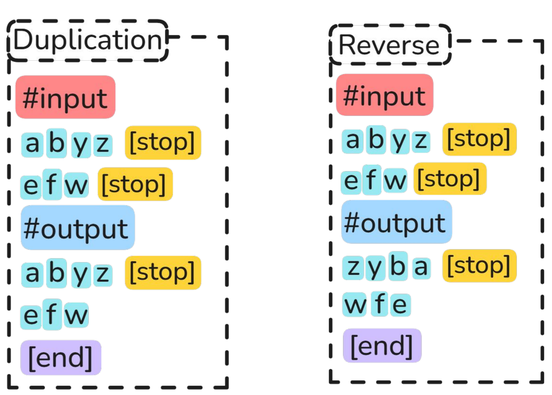


Figure 4 - Duplication and Reverse tasks with multi-inputs

We also propose two new synthetic tasks that were not considered:

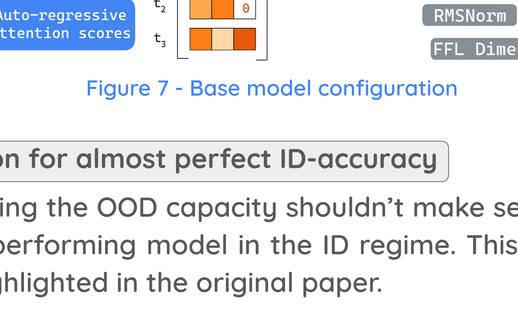


Figure 5 - Our proposed Concatenation and Interleaving tasks

III. Experimental Setup

We opt for the following ID-OOD configuration for our data:

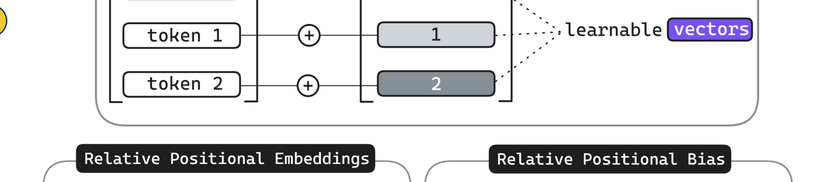


Figure 6 - ID-OOD configuration

We target the GPT architecture, and we first identify a base model configuration capable of achieving almost perfect ID-accuracy (>99% over 10K test samples):

Figure 7 - Base model configuration

Motivation for almost perfect ID-accuracy

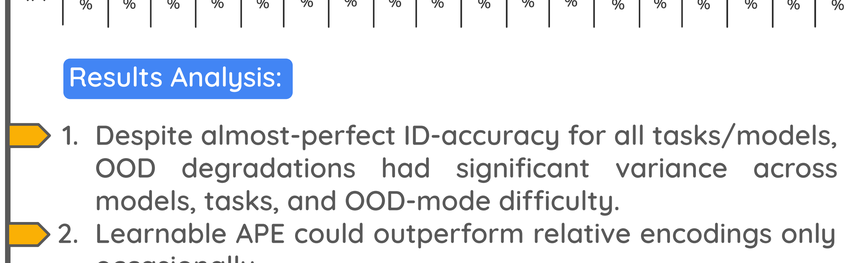
- Assessing the OOD capacity shouldn't make sense for an underperforming model in the ID regime. This point was not highlighted in the original paper.

We compare three different learnable positional encodings for out GPT:

Figure 5 - Tested learnable positional encodings

Beyond architecture: Mimicking Encoder-Decoder with Auto-Regressive Input-Masking

- Encoder-Decoder (ED) models have a natural inductive bias that separates input (which lays in the encoder) from output (which is generated by the decoder).
- Auto-regressive models lack this capacity: they only reason on tokens → may hinder the generalization.
- We propose to implement an Auto-Regressive Input-Masking which mimics ED input-output separation by:
  a. **Limiting the computation of the cross-entropy loss to the output tokens only**
  b. **Training on example chunks instead of random string chunks.**



IV. Results and Conclusion

Concatenation

	In-Distribution			More-Inputs			One-Longer-In			All-Longer-In			More-Inputs + One-Longer-In			More-Inputs + All-Longer-In		
	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB
NTP	99.99 %	100 %	99.99 %	60.78 %	89.91 %	50.92 %	92.69 %	99.92 %	99.40 %	40.94 %	81.48 %	74.30 %	33.93 %	73.32 %	38.64 %	0.00 %	11.34 %	0.45 %
IM	100 %	100 %	100 %	75.12 %	98.71 %	97.90 %	99.48 %	99.98 %	99.99 %	49.42 %	98.54 %	85.37 %	47.00 %	97.43 %	92.81 %	0.00 %	86.82 %	2.78 %

Interleaving

	In-Distribution			More-Inputs			One-Longer-In			All-Longer-In			More-Inputs + One-Longer-In			More-Inputs + All-Longer-In		
	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB
NTP	99.97 %	100 %	100 %	12.56 %	0.00 %	0.44 %				0.00 %	0.00 %	0.00 %				0.04 %	0.00 %	0.00 %
IM	100 %	100 %	100 %	0.00 %	0.00 %	0.00 %				0.00 %	0.00 %	0.00 %				0.00 %	0.00 %	0.00 %

Reversing

	In-Distribution			More-Inputs			One-Longer-In			All-Longer-In			More-Inputs + One-Longer-In			More-Inputs + All-Longer-In		
	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB
NTP	99.85 %	100 %	99.99 %	0.17 %	8.04 %	16.46 %	6.42 %	17.71 %	0.01 %	0.23 %	0.69 %	0.00 %	0.00 %	0.33 %	0.00 %	0.00 %	0.00 %	0.00 %
IM	100 %	100 %	100 %	9.92 %	0.00 %	18.40 %	9.20 %	8.52 %	0.00 %	0.03 %	0.23 %	0.00 %	0.70 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %

Duplicating

	In-Distribution			More-Inputs			One-Longer-In			All-Longer-In			More-Inputs + One-Longer-In			More-Inputs + All-Longer-In		
	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB	APE	RPE	RPB
NTP	99.97 %	100 %	99.99 %	14.36 %	6.34 %	16.84 %	13.60 %	15.37 %	0.00 %	0.94 %	0.00 %	0.00 %	1.86 %	0.06 %	0.00 %	0.00 %	0.00 %	0.00 %
IM	100 %	100 %	100 %	22.95 %	25.03 %	21.18 %	26.76 %	82.19 %	0.00 %	1.62 %	30.78 %	0.00 %	5.89 %	8.62 %	0.00 %	0.00 %	0.00 %	0.00 %

Results Analysis:

1. Despite almost-perfect ID-accuracy for all tasks/models, OOD degradations had significant variance across models, tasks, and OOD-mode difficulty.
2. Learnable APE could outperform relative encodings only occasionally.
3. Input-Masking does not degrade ID-accuracy and dramatically improves OOD but only when base OOD accuracy isn't too low (otherwise it may backfire).

Limitations and Future Work:

1. Limited number of positional encoding techniques were tested → explore the many more remaining.
2. Input-Masking is more supervised → explore architectures that can discover higher level concepts such as input/output in a more self-supervised way.