Controlled tasks for model analysis: Retrieving discrete information fromsequences

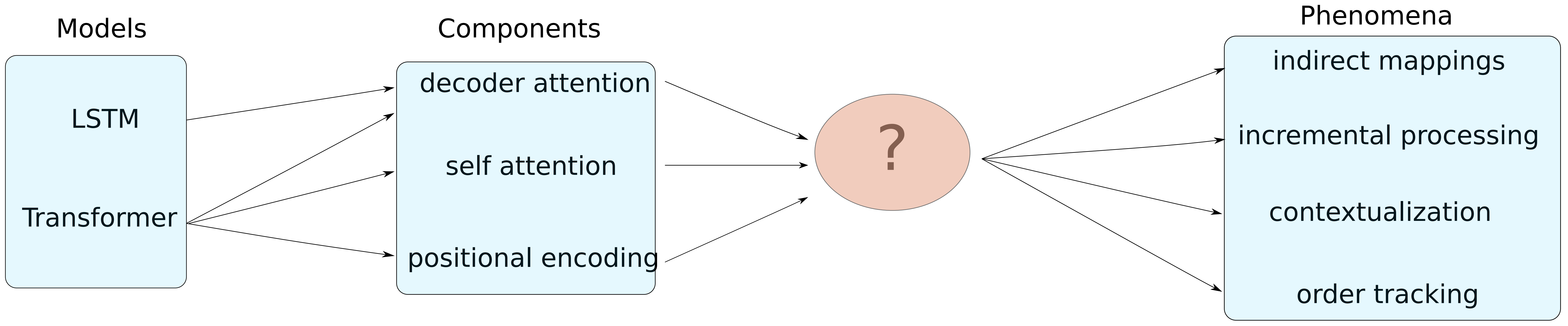
Ionut-Teodor Sorodoc[1], Gemma Boleda[12], Marco Baroni[12]

[1,] Universitat Pompeu Fabra, Barcelona
[2] ICREA

ionut.sorodoc@gmail.com | sorodoc.github.io

Goal

Test the impact of different model components on capturing 4 linguistic phenomena in a controlled environment:


Models

LSTM

Transformer

Components

decoder attention

self attention

positional encoding

?

Phenomena

indirect mappings

incremental processing

contextualization

order tracking

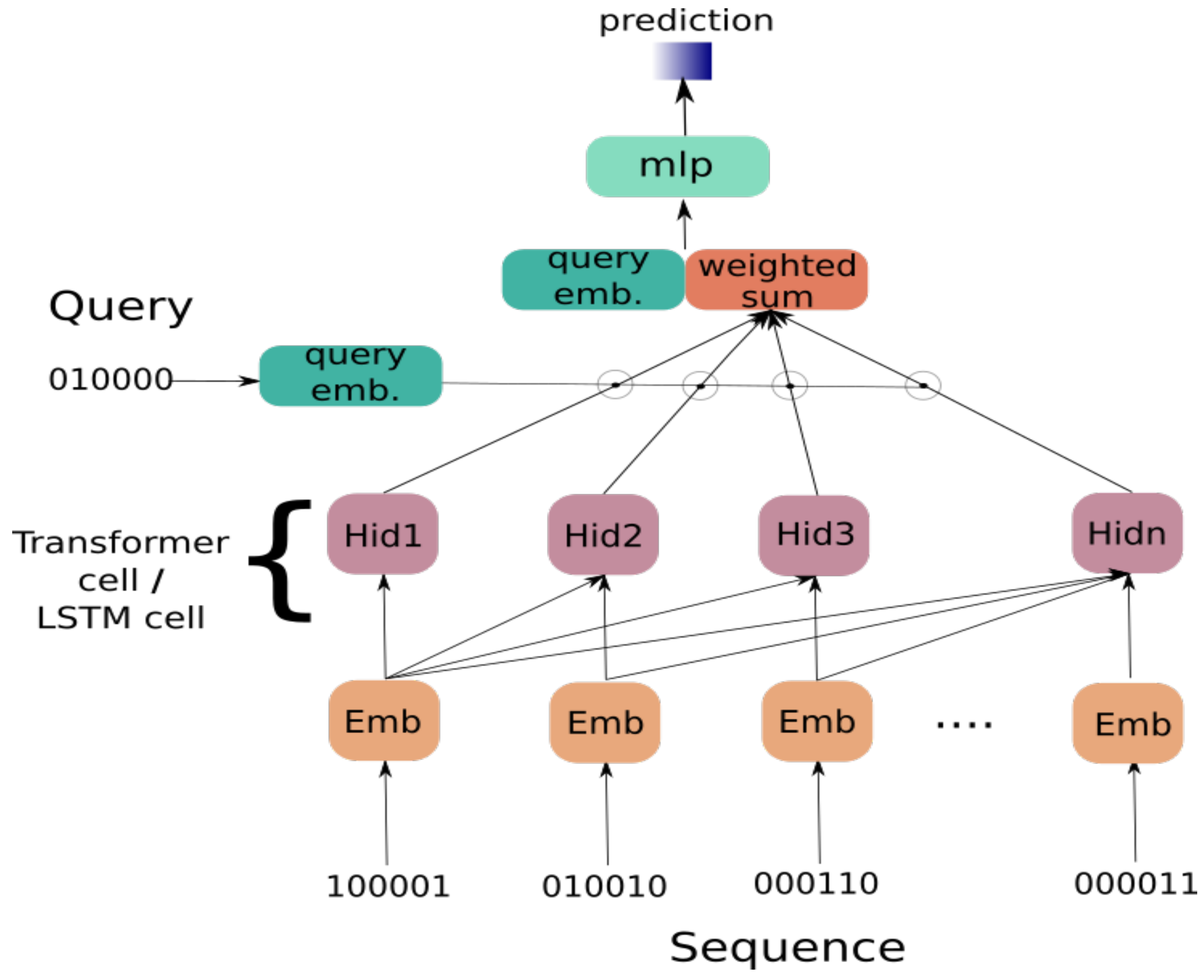
Tasks

input: 100001 010010 000110 ··· 000011

query: 010000

- **T1**: is 2nd feature active in the input?
- **T2**: are 1st or 5th features active in the input? (q: 100010)
- **T3**: are 1st or 5th features active in the input?
- **T4**: are 1st and 5th features active in the input?
- **T5**: are *(1st and 5th)* or *(2nd and 3rd)* active?
- **T6**: does 1st feature appear before 5th feature?
- **T7**: are *(1st before 5th)* or *(2nd before 3rd)* active?

Models


prediction

mlp

query emb.

weighted sum

Query

010000

query emb.

Transformer cell / LSTM cell

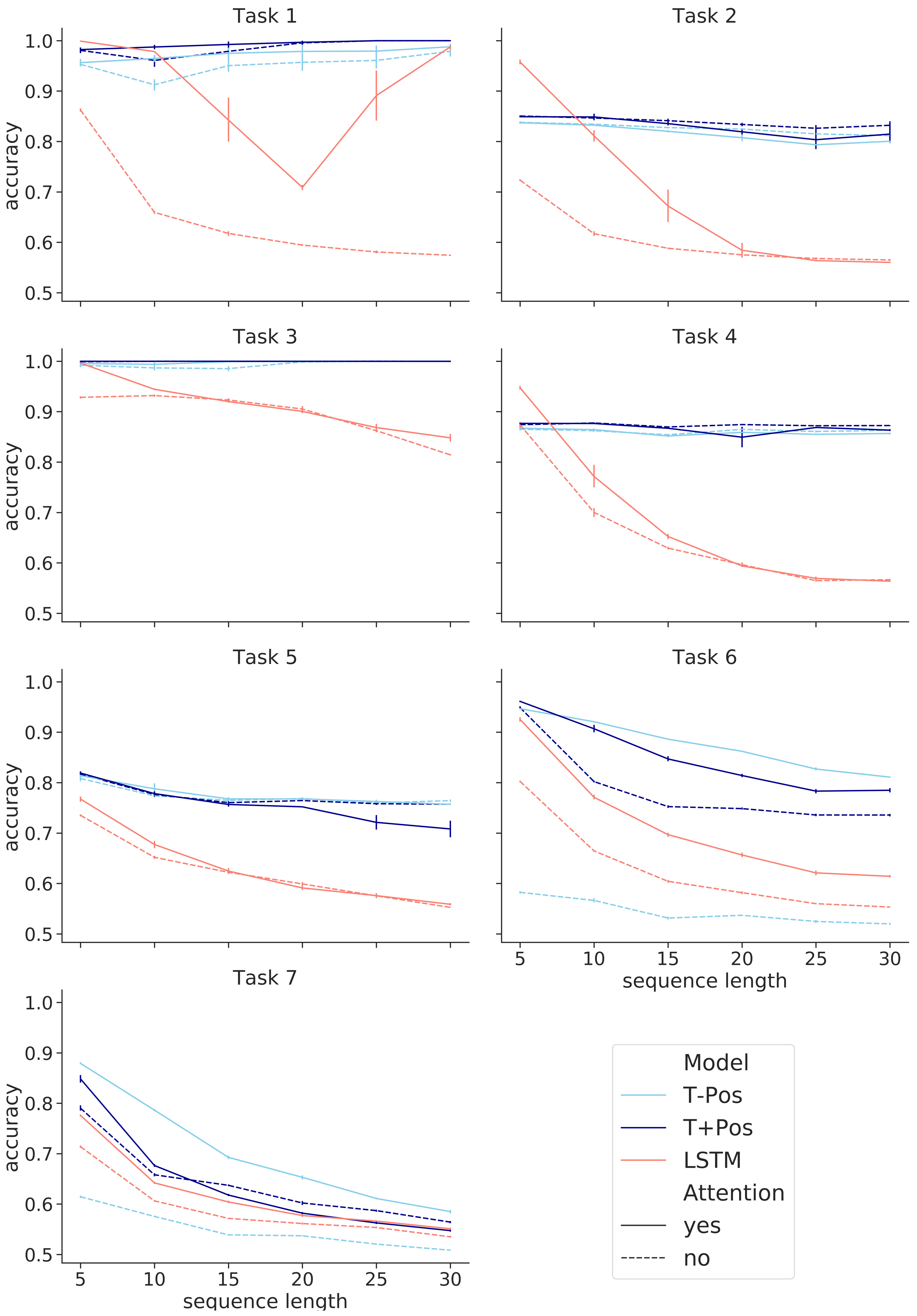
Hid1 Hid2 Hid3 Hidn

Emb Emb Emb .... Emb

100001 010010 000110 000011

Sequence

Results



Conclusions

Transformer better than LSTM, especially for longer sequences

Self attn. and positional encoding behave as noising mechanism in simple tasks

Self attn. is more effective than positional encoding in order tracking

Decoder attention beneficial for single feature extraction

Self attn. not beneficial for T4, T5. Models settle for a degenerate strategy

Self attn. and positional encoding get in the way of each other when learning order tracking