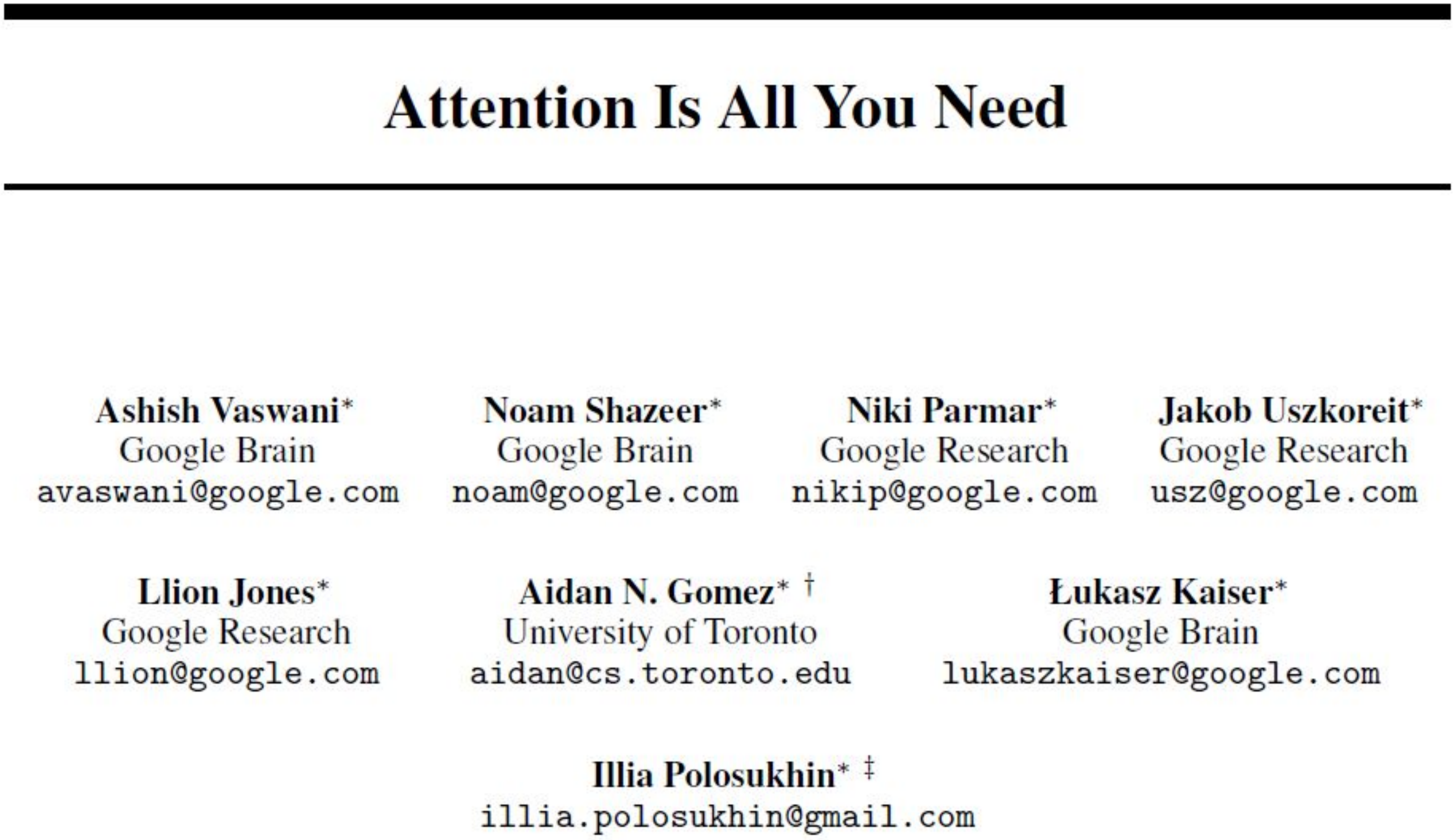
Cognitive Diversity in LLMs Under Memory Constraints

Shayan Shahrabi (shayanshahrabi.github.io),
Saeed Reza Kheradpisheh

Shahid Beheshti University (SBU)

December 2025

Introduction



Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Attention is All you need, Ashish Vaswani, et al. arXiv link

Transformers Are Powerful …

But How Do They Manage Memory?

Working Memory

essential for reasoning, planning, long-context tasks, etc

Different Architecture Behaviors Under Memory Constraints?

Limitation of Previous Work

Next Token Prediction

Linguistic Benchmarks

Proposed Idea

Digit Span

n-back

Models Used

GPT-2 (causal)

GPT-Neo by EleutherAI (causal)

Phi-2 (causal, large)

DistilBERT (masked)

Metrics

Accuray

Memory Capacity

Attention Entropy

$$Corr(entr.\,,\,acc.)$$

Results

DistilBERT = 97.5 %

GPT-2, GPT-Neo < 10 %

Phi-2 ≃ 0 %

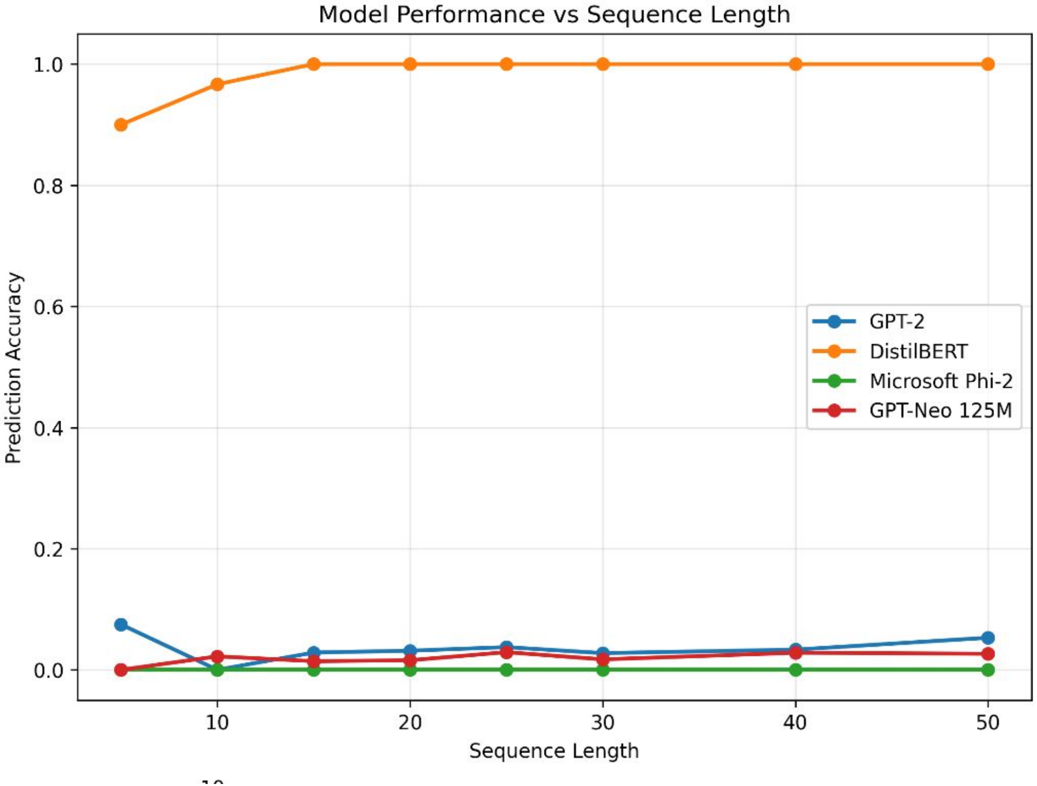
Cognitive Strategies

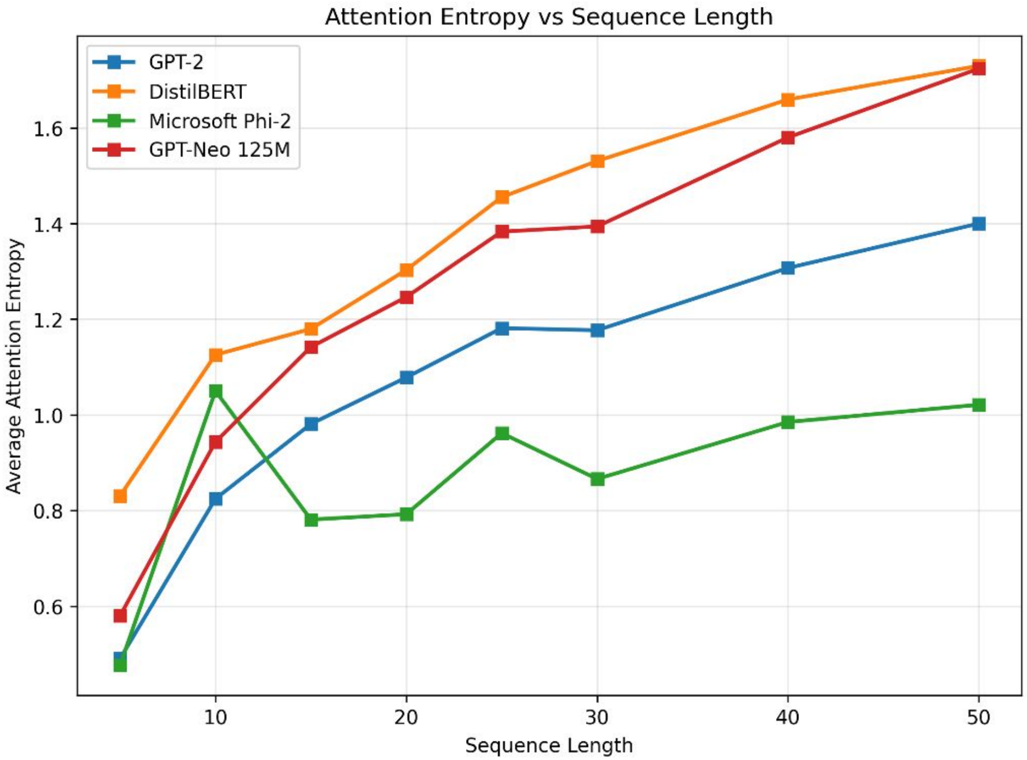
GPT-2 : Increased Entropy

DistilBERT: Decreased Entropy (more focus)

GPT-Neo: Similar to DistilBERT

Phi-2: Unstable, No Clear Strategy


Model Performance vs Sequence Length
Prediction Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
Sequence Length
GPT-2
DistilBERT
Microsoft Phi-2
GPT-Neo 125M


Attention Entropy vs Sequence Length
GPT-2
DistilBERT
Microsoft Phi-2
GPT-Neo 125M
Average Attention Entropy
1.6
1.4
1.2
1.0
0.8
0.6
10
20
30
40
50
Sequence Length

Implications

Model Size

Use of Attention Entropy

Distinct Cognitive Strategies

Choosing Models for Memory Heavy Tasks

Cognitive-inspired AI Design

Interpretability

Questions