BUN-3: Evaluating Transformer Theory of Mind through Child Cognitive Development

Sahasra Kalakonda (sahasrakalakonda16@gmail.com)[1]
[1]Coppell High School

Transformer models provide a computational framework for investigating Theory of Mind (ToM), particularly recursive belief reasoning [2, 7]. ToM refers to the cognitive ability to attribute mental states such as beliefs, intentions, and knowledge to oneself and others [1]. In human development, ToM emerges gradually: children typically fail false-belief tasks at age 3 but succeed around age 4 [3, 8]. More complex second-order reasoning (e.g., “I think that you think…”) develops later, reflecting increased cognitive load [4, 5, 9].

This paper introduces BUN-3, a three-layered benchmark of increasing cognitive complexity: belief retention under falsified evidence (Layer 1), corresponding to first-order false-belief reasoning (reasoning about another agent’s belief) [3, 8]; multi-agent coordination (Layer 2), requiring comparison across agents’ beliefs [4]; and nested second-order belief attribution (Layer 3), involving recursive reasoning (e.g., “I know you know what I know…”) [5, 9]. We extend the OpenToM dataset and evaluate GPT-4o Mini, GPT-3.5 Turbo, and GPT-3 (davinci-002) [6, 7].

GPT-4o Mini achieves near-perfect Layer 1 performance (F1 = 0.96), consistent with robust first-order belief tracking, but declines to 0.90 (Layer 2) and 0.81 (Layer 3), aligning with performance patterns observed in 6-year-olds acquiring second-order reasoning [4, 5]. GPT-3.5 Turbo shows moderate Layer 1 success (F1 = 0.90) but degrades at higher orders (Layer 3: F1 = 0.80), resembling the competence of 4-5-year-olds [3, 4]. GPT-3 performs worst (Layer 1: F1 = 0.77, Layer 3: F1 = 0.31), consistent with pre-competent stages [3, 8]. T-tests and regression analysis confirm that Layer 2 requires significantly more processing than Layer 1 ($t = -2.95$, $p = 0.005$).

BUN-3 provides a psychologically grounded evaluation framework linking transformer reasoning performance to child cognitive development, revealing structured ToM limitations [4, 9]. These findings motivate curriculum-based approaches for improving nested reasoning and establishing a bridge between computational modeling and developmental psychology.

References

[1] D. Premack and G. Woodruff. Does *the chimpanzee have a theory of mind?* 1978.

[2] C. L. Baker et al. *Rational quantitative attribution of beliefs...* 2017.

[3] J. Perner, S. R. Leekam, and H. Wimmer. *Three-year-olds' difficulty with false belief.* 1987.

[4] H. M. Wellman and D. Liu. *Scaling of theory-of-mind tasks.* 2001.

[5] S. Grueneisen et al. *Second-order false-belief reasoning.* 2015.

[6] Y. Xu et al. *OpenToM benchmark.* 2024.

[7] OpenAI. *GPT-4o Mini technical report.* 2024.

[8] H. Wimmer and J. Perner. *Beliefs about beliefs (false-belief task).* 1983.

[9] I. A. Apperly. *Mindreaders: The Cognitive Basis of "Theory of Mind."* 2010.

Figures and Tables

Table 1-3: Results Table

TABLE I

LAYER 1: MACRO F1 SCORE (0–1, HIGHER = BETTER ACCURACY) OF GPT-4O MINI, GPT-3.5 TURBO, AND GPT-3 (DAVINCI-002) EVALUATED ON THE EXTENDED OPENTOM USING COT PROMPTING (CHAIN-OF-THOUGHT), ALONG WITH AVERAGE RESPONSE TIME IN SECONDS. RELATIVE AGE OF CHILD RECORDED ACCORDING TO CLASSIC CHILD PSYCHOLOGY EXPERIMENTS. RELEVANT PERFORMANCES ARE SHOWN AS AGE ANSWERS WITH HIGH ACCURACY , AGE ANSWERS WITH MODERATE ACCURACY , OR AGE ANSWERS WITH LOW ACCURACY .

Question	GPT-4o Mini		GPT-3.5 Turbo		GPT-3 (davinci-002)	
$\text{Loc}_c(F)$	0.98	0.66	0.87	1.21	0.77	1.77
$\text{Loc}_c(S)$	1.00	0.59	0.79	1.21	0.75	0.89
$\text{Loc}_f(F)$	1.00	0.69	0.96	1.06	0.81	0.82
$\text{Loc}_f(S)$	0.88	0.65	1.00	0.73	0.75	1.23
MHop(F)	1.00	0.61	0.85	0.78	0.76	0.85
MHop(S)	0.98	0.56	0.83	0.50	0.70	0.78
Att	0.88	0.54	1.00	0.56	0.88	0.80
Total Avg	0.96	0.61	0.90	0.86	0.77	1.02

Cognitive Child Age
3.0 years old
3.5 years old
4.0 years old
4.5 years old
5 years old
5.5 years old
6 years old

TABLE II

LAYER 2: MACRO F1 SCORE (0–1, HIGHER = BETTER ACCURACY) OF GPT-4O MINI, GPT-3.5 TURBO, AND GPT-3 (DAVINCI-002) EVALUATED ON THE EXTENDED OPENTOM USING COT PROMPTING (CHAIN-OF-THOUGHT), ALONG WITH AVERAGE RESPONSE TIME IN SECONDS. RELATIVE AGE OF CHILD RECORDED ACCORDING TO CLASSIC CHILD PSYCHOLOGY EXPERIMENTS. RELEVANT PERFORMANCES ARE SHOWN AS AGE ANSWERS WITH HIGH ACCURACY , AGE ANSWERS WITH MODERATE ACCURACY , OR AGE ANSWERS WITH LOW ACCURACY .

Question	GPT-4o Mini		GPT-3.5 Turbo		GPT-3 (davinci-002)	
$\text{Loc}_c(F)$	0.95	0.65	0.89	1.31	0.34	1.69
$\text{Loc}_c(S)$	0.97	0.69	0.99	1.48	0.75	1.74
$\text{Loc}_f(F)$	0.84	0.80	0.81	1.21	0.80	1.69
$\text{Loc}_f(S)$	0.90	1.02	0.79	1.20	0.58	1.71
MHop(F)	0.88	0.93	0.79	1.26	0.73	1.66
MHop(S)	0.87	1.04	0.92	1.34	0.38	1.81
Att	0.88	0.68	0.83	1.21	0.81	1.73
Total Avg	0.90	0.83	0.86	1.29	0.63	1.73

Cognitive Child Age
3.0 years old
3.5 years old
4.0 years old
4.5 years old
5 years old
5.5 years old
6 years old

TABLE III

LAYER 3: MACRO F1 SCORE OF GPT-4O MINI, GPT-3.5 TURBO, AND GPT-3 (DAVINCI-002) EVALUATED ON THE EXTENDED OPENTOM USING COT PROMPTING (CHAIN-OF-THOUGHT) AND RELATIVE AGE OF CHILD RECORDED ACCORDING TO CLASSIC CHILD PSYCHOLOGY EXPERIMENTS. THE RELEVANT PERFORMANCES ARE SHOWN AS AGE ANSWERS WITH HIGH ACCURACY , AGE ANSWERS WITH MODERATE ACCURACY , OR AGE ANSWERS WITH LOW ACCURACY .

Question	GPT-4o Mini		GPT-3.5 Turbo		GPT-3 (davinci-002)	
$\text{Loc}_c(F)$	0.79	1.41	0.77	1.65	0.10	2.48
$\text{Loc}_c(S)$	0.87	2.46	0.75	2.90	0.45	2.91
$\text{Loc}_f(F)$	0.89	1.76	0.87	1.96	0.16	2.43
$\text{Loc}_f(S)$	0.61	1.54	0.68	1.82	0.26	2.42
MHop(F)	0.77	1.63	0.92	2.21	0.45	2.42
MHop(S)	0.92	1.55	0.84	1.81	0.13	2.46
Att	0.85	1.99	0.78	1.64	0.64	2.39
Total Avg	0.81	1.76	0.80	2.00	0.31	2.50

Cognitive Child Age
3.0 years old
3.5 years old
4.0 years old
4.5 years old
5 years old
5.5 years old
6 years old

Figure 1-3: Response Time Linear Regression

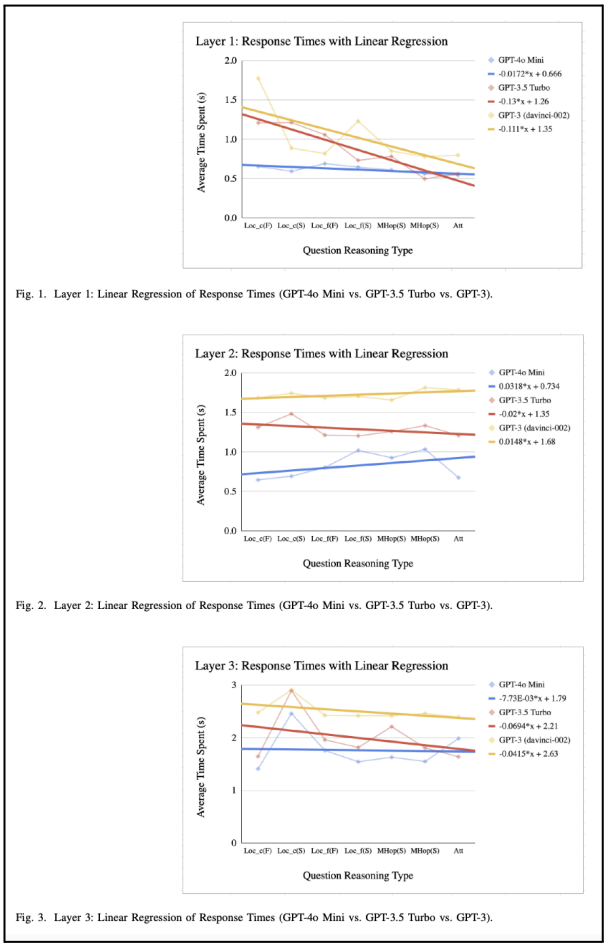

Fig. 1. Layer 1: Linear Regression of Response Times (GPT-4o Mini vs. GPT-3.5 Turbo vs. GPT-3).

Fig. 2. Layer 2: Linear Regression of Response Times (GPT-4o Mini vs. GPT-3.5 Turbo vs. GPT-3).

Fig. 3. Layer 3: Linear Regression of Response Times (GPT-4o Mini vs. GPT-3.5 Turbo vs. GPT-3).