1. Extra Experiment Results for Response to Reviewer gFV2

A1 for Q1.

Table 1: We conducted additional experiments and calculated the percentages for GPT-2, GPT-2-Large, and GPT-2-XL across different datasets. The results show that larger GPT models achieve similar percentages to BERT, e.g., around 98%.

Model	gsm8k	Yelp	GLUE	DailyMail	OpenOrca	WikiText	Avg. Percentage
GPT-2 (124M)	75.19	77.46	77.49	73.11	69.32	72.31	75.15
GPT-2-Large (774M)	98.49	98.47	98.16	98.17	98.34	98.08	98.29
GPT-2-XL (1.5B)	98.64	98.32	97.85	97.83	97.90	97.80	98.05

A2 for Q2.

Table 2: We used two popular dependency parsing methods: Stanza (Stanford NLP) and AllenNLP. The results for verifying truthful semantic dependencies encoded in the final layers are similar to those obtained with SpaCy.

	BERT	RoBERTa	tinyRoBERTa	ALBERT	DistilBERT	DeBERTa	MobileBERT	MiniLM	GPT-2	LLaMA3
SpaCy	87.86	87.71	82.44	88.77	88.88	87.17	85.8	84.62	93.41	92.47
Stanza	84.33	86.9	81.14	85.53	87.19	83.69	80.98	83.67	91.42	90.32
AllenNLP	83.02	84.04	80.23	84.32	85.54	82.98	81.54	79.87	90.35	90.25

A4 for Q4.

Table 3: Additional experiments using more advanced ChatGPT-4o model to compare the model's answer with the ground truth and find incorrect cases. The results are similar with using F1<0.6.

	BERT	RoBERTa	tinyRoBERTa	ALBERT	DistilBERT	DeBERTa	MobileBERT	MiniLM	GPT-2	LLaMA3
p (F1<0.6)	79.07	69.2	77.94	71.86	81.8	75.32	66.61	77.56	48.04	64.56
F1	92.93	84.86	82.83	80.56	85.71	91.69	81.19	85.34	0.78	35.81
p (GPT-4o select)	79	68.42	73.31	66.11	81.79	77.84	68.69	77.56	59.6	62.35
accuracy	88.45	78	78	74.63	76.63	90.44	74.5	78.9	0.1	14.68

A5 for Q5.

Table 4: Extra experiments using a one-shot setting, which aligns with official benchmark evaluations.

Model	F1 (0-shot)	F1 (1-shot)
GPT-2 (124M)	0.78	5.5
GPT-2-Large (774M)	7.3	21.09
LLaMA3-8B-instruct (8B)	35.81	76.27

2. Extra Experiment Results for Response to Reviewer LWrx

A1 for Q1.

Table 5: Additional experiments using 10 independent random samples per token. The results remained very similar, further validating the stability of our results.

	BERT	RoBERTa	tinyRoBERTa	ALBERT	DistilBERT	DeBERTa	MobileBERT	MiniLM	GPT-2	LLaMA3
$k = 5$	98.81	93.06	94.29	97.01	95.11	99.62	96.49	88.69	75.15	95.59
$k = 10$	98.72	93.21	93.98	97.22	94.83	95.46	95.32	86.95	76.32	95.83

3. Extra Experiment Results for Response to Reviewer VaRV

A1 for Q1.

Table 6: We conducted additional experiments and calculated the percentages for GPT-2, GPT-2-Large, and GPT-2-XL across different datasets. It suggests that semantic retention is also influenced by other factors such as model complexity.

Model	gsm8k	Yelp	GLUE	DailyMail	OpenOrca	WikiText	Avg. Percentage
GPT-2 (124M)	75.19	77.46	77.49	73.11	69.32	72.31	75.15
GPT-2-Large (774M)	98.49	98.47	98.16	98.17	98.34	98.08	98.29
GPT-2-XL (1.5B)	98.64	98.32	97.85	97.83	97.90	97.80	98.05

A2 for Q2.

Table 7: The two-by-two possibility table for model answer correctness and semantic dependency correctness. $P(f_\theta)$ stands for the percentage when the model answers correctly and semantic dependency is correctly encoded. $P'(f_\theta)$ stands for the percentage when the model answers incorrectly and semantic dependency is incorrectly encoded.

	Correct Semantic Dependency	**Incorrect Semantic Dependency**
Model Answer Correctly	$P(f_\theta)$	$1 - P(f_\theta)$
Model Answer Incorrectly	$1 - P'(f_\theta)$	$P'(f_\theta)$

Table 8: All four percentages for all models. The results show that when the model correctly encodes the semantic dependency in the final-layer token, it usually provides the correct answer. Conversely, when the model produces an incorrect answer, the semantic dependency is often incorrectly encoded. These findings highlight the importance of semantic dependency encoded in the final-layer token for model predictions.

	BERT	RoBERTa	tinyRoBERTa	ALBERT	DistilBERT	DeBERTa	MobileBERT	MiniLM	GPT-2	LLaMA3
$P(f_\theta)$	93.26	82.32	83.33	87.05	96.48	89.25	75.24	91.97	81.25	70.56
$1 - P(f_\theta)$	6.74	17.68	16.67	12.95	3.52	10.75	24.76	8.03	18.75	29.44
$P'(f_\theta)$	79.07	69.2	77.94	71.86	81.8	75.32	66.61	77.56	48.04	64.56
$1 - P'(f_\theta)$	20.93	30.8	22.06	28.14	18.2	24.68	33.39	22.44	51.9	35.44

A6 for Q6.

Table 9: We provide extra major experiments on recent open-source models like the Qwen model. In the future, we will test more new models such as Deepseek, Phi-4, and Mistral. Exp1.Two validations on basic mechanisms of token-level semantic information propagation.

Models	Validation 1 (Self-Information Retention)	Validation 2 (Sequence-Level Semantic Aggregation)
BERT	98.81	99.29
Qwen2-1.5B-Instruct (new)	96.91	100.00

Table 10: Exp2.Alignment score that indicates how well individual tokens encode truthful semantic dependencies.

Models	Average Alignment Score (%)
BERT	87.86
Qwen2-1.5B-Instruct (new)	93.51

Table 11: Exp3.The percentage of failed QA cases matches our semantic dependency assumption.

Models	Percentage P (f_θ) (%)	Average F1 Score (%)
BERT	87.86	92.93
Qwen2-1.5B-Instruct (new)	52.38	24.93

4. Extra Experiment Results for Response to Reviewer 1yCg

A3 for Q3.

Table 12: The two-by-two possibility table for model answer correctness and semantic dependency correctness. $P(f_\theta)$ stands for the percentage when the model answers correctly and semantic dependency is correctly encoded. $P'(f_\theta)$ stands for the percentage when the model answers incorrectly and semantic dependency is incorrectly encoded.

	Correct Semantic Dependency	**Incorrect Semantic Dependency**
Model Answer Correctly	$P(f_\theta)$	$1 - P(f_\theta)$
Model Answer Incorrectly	$1 - P'(f_\theta)$	$P'(f_\theta)$

Table 13: All four percentages for all models. The results show that when the model correctly encodes the semantic dependency in the final-layer token, it usually provides the correct answer. Conversely, when the model produces an incorrect answer, the semantic dependency is often incorrectly encoded. These findings highlight the importance of semantic dependency encoded in the final-layer token for model predictions.

	BERT	RoBERTa	tinyRoBERTa	ALBERT	DistilBERT	DeBERTa	MobileBERT	MiniLM	GPT-2	LLaMA3
$P(f_\theta)$	93.26	82.32	83.33	87.05	96.48	89.25	75.24	91.97	81.25	70.56
$1 - P(f_\theta)$	6.74	17.68	16.67	12.95	3.52	10.75	24.76	8.03	18.75	29.44
$P'(f_\theta)$	79.07	69.2	77.94	71.86	81.8	75.32	66.61	77.56	48.04	64.56
$1 - P'(f_\theta)$	20.93	30.8	22.06	28.14	18.2	24.68	33.39	22.44	51.9	35.44

A4 for Q4.

Table 14: Extra experiments using a one-shot setting, which aligns with official benchmark evaluations.

Model	F1 (0-shot)	F1 (1-shot)
GPT-2 (124M)	0.78	5.5
GPT-2-Large (774M)	7.3	21.09
LLaMA3-8B-instruct (8B)	35.81	76.27