Transformer-based language models and complement coercion: Experimental studies

Yuling Gu
University of Washington, Seattle, WA, USA
yulinggu@uw.edu

1. OVERVIEW

How do **transformer-based language models (LMs) react** to **implicit meaning**?

- **Complement coercion**

  e.g. "*The student finished the book about sailing*" where the action "*read*" is implicit

- Compare **LMs' surprisal estimates** at various critical sentence regions

Condition	Sentence
Coerced	😮 (*LM Surprisal* ↑) The student finished **the book** about learning how to sail.
Preferred	😐 The student read **the book** about learning how to sail.
Non-preferred	😐 The student wrote **the book** about learning how to sail.

Critical sentence regions:

- Differing verb : finished/read/wrote
- Target region: the book
- Post-target region: about learning

Is 😮 because of processing:

- ❑ Implicit meaning
- ❑ Non-preferred condition
- ❑ Entity NP
- ❑ Less specific verb
- ❑ Event interpretation of NP
- ❑ Anomaly detection

.....

Figure 1: Example of a set of test sentences in Experiment 1 and the critical regions for measurement.

2. WHAT IS COERCION?

Environment it can occur in

Verbs like *started, finished, completed* semantically select for an event-describing complement

Default interpretation: entity

Type-mismatch!

Coerced: The student finished the book about learning how to sail.

Uncovering implicit meaning

Step1: Event-selecting verb (semantics)
read, wrote, ate, watched etc

Step2: World-knowledge
read, wrote, ~~ate, watched etc~~

Step3: More world-knowledge
read, ~~wrote, ate, watched etc~~

Semantics + world knowledge

Control: The student read the book about learning how to sail.

3. EXPERIMENT DESIGN

- **Surprisal :** quantification of cognitive effort required to process a word in a sentence

$$S(w_i) = -\log_2 p(w_i|w_1, \cdots, w_{i-1})$$

- **Measure positions :** 3 critical regions (Figure 1)
- **Models :** family of GPT-2 models

- **Diagnostic datasets :**

Dataset	Original psycholinguistic experiment	Selection from original stimuli
1	"Coercion in sentence processing: Evidence from eye-movements and self-paced reading" by Traxler et al. (2002)	36 triplets (Coercion/Preferred/Non-preferred) from stimuli for Experiment 1
2		32 quadruplets (Event/Neutral verb + Event/Entity NP) from stimuli for Experiments 2 and 3
3	"An MEG Study of Silent Meaning" by Pylkkänen and McElree (2007)	35 triplets (Coerced/Anomalous/Control) from the *Nonembedded Stimuli*

Table 1: Source of diagnostic datasets used in our experiments.

4. RESULTS AND ANALYSIS

Experiment 1

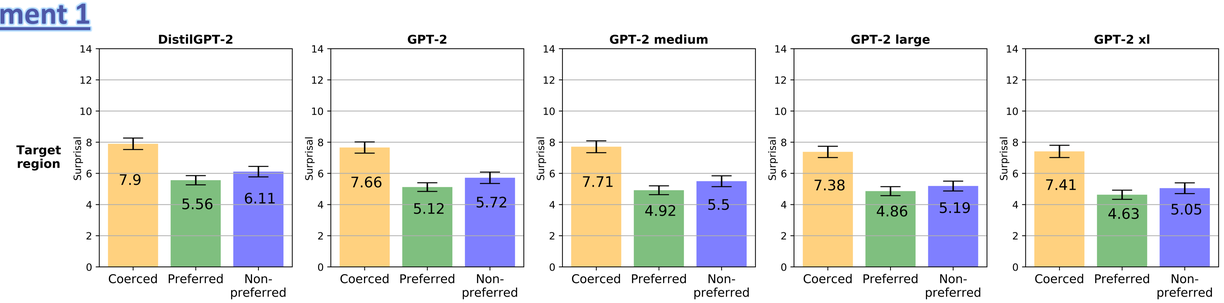

Figure 2: Bar graphs showing mean surprisal estimates from Experiment 1, by model, region and condition. Error bars represent standard error.

At the target region, surprisal in ...

- coerced condition >> preferred condition
- coerced condition >> non-preferred condition
- preferred condition ≈ non-preferred condition

😮 **is because of processing:**

- ☑ **Implicit meaning**
- ☒ **Non-preferred condition**

Experiments 2 & 3

😮 **is because of processing:**

- ☑ **Entity NP**
- ☒ **Less specific verb**
- ☒ **Event interpretation of NP**
- ☑ **Anomaly detection**

5. TAKEAWAYS

❖ Our work is the **first of its kind** to study **transformer-based LMs' behavior** on the **complement coercion phenomenon** using **surprisal estimates**.

❖ While previous works studying LMs' behavior compare full sentences or examine one critical region per sentence, for each sentence, we take **measurements at three positions** important for analysis of the phenomenon to provide a **richer analysis**.

❖ The series of three experiments we perform provide an illustrative example of how **targeted follow-up experiments** could be used to **tease apart confounding factors**.