DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models

By: Marjanović et al.
Seminar: Cross temporal NLP
Lecturer: Wei Zhao

Structure

- Definition of Knowledge Conflicts
- Motivation
- Preparations
- Experiments
- Results
- Discussion Starters

Defining Knowledge Conflicts

- Intra Memory Conflicts
  - Deviating training data
  - Conflicting view points
  - Temporal Changes
- Context Memory Conflict
  - Given context valued less than learned parameters

Motivation

- LLMs train on massive amounts of data:
  - Out of date
  - Meaning incongruent across time

LLMs giving out differing information depending on given context

Researcher’s Intention

- Investigate intra-memory conflicts
- Create a QA dataset focusing on temporally and factually conflicting facts
  - 11,378 question-answer pairs
- Introduce two units for measurements:
  - dynamicity
  - staticness

→ How much can a fact change?

⇒ Create a dataset to observe the effects of intra-memory and context memory conflicts using two new measurements

Units explained

- Static Questions
  - Topics that do not generally change
- Dynamic Questions
  - Topics that change

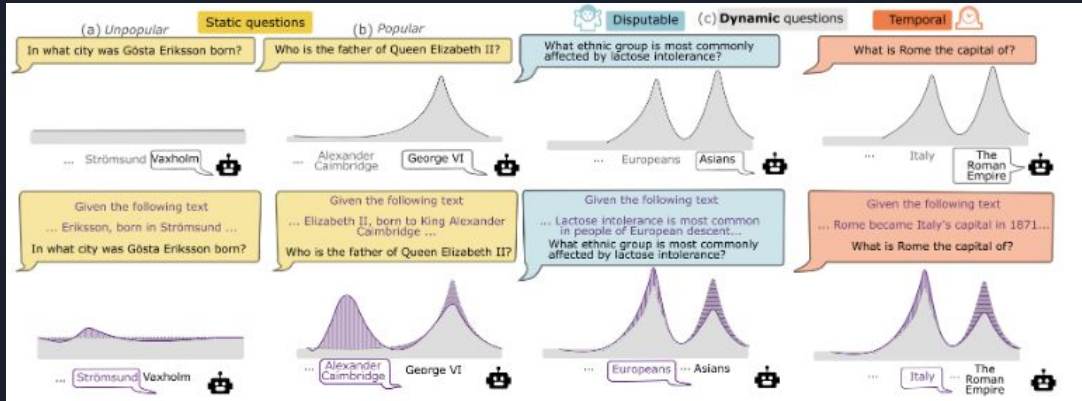

Figure 1: We present examples of our dataset here, consisting of static, temporal, and disputable facts. We show how model output distribution can vary due to the popularity (a,b) and dynamicity (c) of facts. Fact dynamicity (c) causes intra-memory conflicts between the different fact representations seen during pretraining. In the bottom row, we show the change in output probability (purple area) that the context must enact on the initial output distribution (dotted line) to force a new output distribution (purple line).

Question Generation

Two proxy units for measurement:

- Temporality
  - How many Page Views were there? → Popularity Score
- Disputability
  - Taken from controversial Wikipedia posts
  - Number of edits relay degree of Disputability
    - not counting vandalism or trolling

Llama 3 8B Instruct is used to generate the Question

- using context
- providing a ground-truth answer

Question Generation (Llama prompting)

	Prompt P
Input $[P; q_i]$	"System": You'll be given a question and a context about the article and answer it with a one word. Answer the [Question], "User": This article is about Titanic. Who was the producer of Titanic?
Input $[P; c_i; q_i]$	"System": You'll be given a question and a context about the article and answer it with a one word. Answer the [Question], "User": This article is about Titanic. Titanic is a 1997 American epic romantic disaster film directed, written, produced, and co-edited by James Cameron. Incorporating both historical and fictionalized aspects, it is based on accounts of the sinking of RMS Titanic in 1912. Leonardo DiCaprio and Kate Winslet star as members of different social classes who fall in love during the ship's maiden voyage. The film also features an ensemble cast of Billy Zane, Kathy Bates, Frances Fisher, Gloria Stuart, Bernard Hill, Jonathan Hyde, Victor Garber, David Warner, Suzy Amis and Bill Paxton. Who was the producer of Titanic?

Table 4: Example of a prompt according to the presence of the context c_i in the input. The example here is the question that asks about the producer of the movie Titanic with the answer-specific context which mentions answer (James Cameron) in the context.

Question Generation (Disputable Facts)

To identify disputable facts from the controversial articles they do:

- find amount of revisions in an article's logs
- select two logs as a pair
- check if the texts are identical
- measure the edit distance between the two
    - if a pair of words yields the same texts it is chosen

Question Generation (finding vandalism)

Vandalism is suspected when:

- User involved is anonymous
- Has their IP-Adress as their ID

These are not counted

Then:

- check word pairs again for synonyms/paraphrasing
- remove any pairs whose similarity is larger than 0.98 (checked by semantic similarity model)

Question Generation (Annotation)

Undisclosed amount of annotators annotated questions:

- If multiple questions focus same context, most specific is kept
- Annotator is given 5 Options:
  - Accept
  - Change
  - Discard
  - Misinformation
  - Incomplete

Two annotators annotated every instance, in case of disagreements a third annotator was brought in.

Hypothesis Formulation

Highly dynamic information should result in:

- higher entropy
- greater change in output distribution

Measuring (Preliminaries)

They declare that a dataset will consist of instances *i*, where each instance has:

- A context *c*
- A question *q*
- An answer *y*

Which they combine into:

- Input with context == $x_{i;q;c}$
- Input without context == $x_{i;q}$

Measurements (Semantic Entropy)

1. Collect Semantic Sets
    a. Let model generate K outputs
    b. Group answers by semantic similarity (using DeBERTA NLI model)
2. Compute entropy between sets
    a. obtain conditional token probability from the generating model
    b. approximate overall semantic entropy of a group using Monte Carlo Integration

Semantic Entropy Formula

p is the conditional token probability given g (answer) and x (input)

h refers to the intermediate tokens in the entire output length

V is the amount of groups G available

$$p(g_v|x_i) = \sum_{y_{i,k} \in g_v} p(y_{i,k}|x_i)$$
$$= \sum_{y_{i,k} \in g_v} \prod_h p(y_i^h | y_i^{<h}, x_i)$$

$$SE(x_i) \approx -V^{-1} \sum_{v=1}^{V} \log p(g_v|x_i)$$

Measurements (Coherent Persuasion Score)

A score to measure the actual efficacy of a provided context to changing a model's output

- Previous persuasion scores only focused on first token of a single answer

Instead, they provide a novel approach to this score:

- multi-sample approach → grouping semantically similar answers
- creating answer distributions for context missing and available

Measurements (Coherent Persuasion Score)

To calculate, they gather:

- two lists of answers (with or without context)
- create semantic similarity groups
- average the divergence between both lists

They pose that it is “more coherent” due to:

- it considering the entirety of an LM’s output
- only capturing the semantic divergence of an output

CP Score Formula

K is amount of model outputs

R is amount of semantically similar no context groups

U is amount of semantically similar context groups

y is one answer from the answer lists

W is $\in \{R, U\}$

p_{yw} is the averaged softmax probability distribution

$$p_{g_w} = \frac{1}{W} \sum_{w=1}^{W} p_{y_w} \quad (3)$$

$$CP(c_i) = \frac{1}{|R| \times |U|} \sum_{r=1}^{R} \sum_{u=1}^{U} KL(p_{g_r}, p_{g_u}) \quad (4)$$

Experimental Setup

Use three instruction tuned models:

- Mistral 7B Instruct v0.1
- Llama 2 7b chat hf
- Qwen2 7B Instruct

In a zero shot setting, they query a model:

- first without context
- second with two types of context:
  - unperturbed context
  - unseen replacement

Experimental Setup

- They collect accuracy and semantic entropy per query
- calculate CP Score
- Each query gives out 10 samples
  - greedy search takes the most likely example for accuracy calculation

$$acc = \frac{\sum_{i=1}^{N} RougeL(a_i, y_i) > 0.3}{N}$$

Analysis

Identify the difficulty of updating dynamic facts through two behaviours:

- Persuaded
  - model got convinced through context
- Stubborn
  - model did not get convinced

Compare obtained CP and SE scores across full dataset

Investigate Persuasion instances through computing Pearson Correlation between:

- semantic entropy
- temporality
- popularity

Results

Static facts are the easiest for models to accept context on.

Temporal and Disputable facts are similarly difficult to trust for models.

Overall, Llama is more easily persuaded.

	# of Questions	# of Instances	% of Stubborn Instances			% of Persuaded Instances		
			Llama-2	Mistral	Qwen2	Llama-2	Mistral	Qwen2
Static	2500	5000	6.16%	5.44%	6.92%	**78.44%**	**70.52%**	**61.48%**
Temporal	2495	4990	**9.38%**	**7.01%**	**7.54%**	60.96%	51.62%	44.81%
Disputable	694	1388	9.36%	6.48%	7.35%	63.83%	62.53%	59.51%

Table 1: The number of collected questions and instances in DYNAMICQA (§3) for each fact type. We also report general model behaviour (i.e. percentage of persuaded instances given context), as further described in §5.2.

Results

Lower SE can be interpreted as more consistent and less conflicted output.

CP score demonstrates efficacy of context shifting output distribution.

		Accuracy (↑)			Semantic Entropy (↓)			Coherent Persuasion Score (↑)		
		Llama-2	**Mistral**	**Qwen2**	**Llama-2**	**Mistral**	**Qwen2**	**Llama-2**	**Mistral**	**Qwen2**
Static	with context	**0.8476**	**0.7644**	**0.6814**	15.5663	11.7557	10.5875	**6.8665**	**5.8550**	**4.1567**
	w/o context	**0.1306**	**0.0902**	**0.1244**	17.7064	**11.4943**	10.4271			
Temporal	with context	0.6619	0.5677	0.5040	**15.3947**	**10.7685**	**10.5264**	6.5941	5.6314	3.9551
	w/o context	0.1036	0.0719	0.0866	**17.3518**	11.7410	11.0875			
Disputable	with context	0.6455	0.6253	0.5937	16.5803	11.6632	10.9627	5.6027	4.1147	3.3955
	w/o context	-	-	-	18.9694	12.4214	**10.3957**			

Table 2: The average accuracy, Semantic Entropy (SE; §4.2) and Coherent Persuasion (CP; §4.3) score of our models. We bold the best values per column, with and without context. Given the inherent subjectivity of the Disputable facts, we do not show accuracy without context.

Obstacles to persuasion

Dynamic (temporal) facts show a higher proportion of stubborn instances but a smaller divergence in output distribution.

- calculate loss over all tokens in the target answer
  - temporal and disputable questions exhibit greater losses
- Significance is confirmed via Welch's t-test

It takes more effort to update a LM's parameters for dynamic facts

Interacting with Persuasion

Stubborn instances have lower CP scores, but not necessarily differing SE scores compared to Persuaded or Neutral instances.

→ High Initial entropy does not necessarily result in persuasion.

While popularity negatively impacts persuasion, popular instances are not over-represented in stubborn instances.
Whereas temporality scores show the strongest correlation with CP.

→ Persuasion and temporality are stronger connected than persuasion and popularity or entropy

Conclusion

Static facts are the easiest to update with context.

Number of unique presentations of a fact on Wikidata/Wikipedia has an inverse correlation to a model's chance for update adoption.

Depending on the LM, persuasion chance also varies.

- Llama is easier persuaded than the others.

Uncertainty does not necessarily improve persuasion rate.

Discussion Starters

Do you see any reason for why the three models are differing to such an extent in persuasion chance?

Would it always be better if a model is persuaded by a given context?