*Under review as a conference paper at *the 1st Conference on Language Modeling* (**COLM 2024**)

*Under review as a workshop paper at *the 41st International Conference on Machine Learning* (**ICML 2024 – Workshop on Mechanistic Interpretability**)

Unveiling LLMs: The Evolution of Latent Representations in a Temporal Knowledge Graph*

Marco Bronzini, *Carlo Nicolini*, *Bruno Lepri*, *Jacopo Staiano*, and *Andrea Passerini*

UNIVERSITÀ DI TRENTO

Ipazia

FONDAZIONE BRUNO KESSLER

#LLMs #interpretability #latentRepresentation #knowledge #graphs

About me

Marco Bronzini

- PhD student at the University of Trento (Italy)
- Focused on: #LLMs, #Interpretability, #Sustainability, #ESG

ACADEMIC BACKGROUND:

- Bachelor's degree in *"Computer Science for Management"* at the University of Bologna (Italy)
- Master's degree in *"Data Science"* at the University of Trento (Italy)

WHY AM I HERE?

- one-month visiting period at *INRIA Paris* via a joint agreement between *Collège de France* and my own university → supervised by Benoît Sagot
- May 6th – June 7th 2024

UNIVERSITÀ DI TRENTO → COLLÈGE DE FRANCE 1530 → Inria

CURIOUS FACT:

- Met some of you at the *Advanced Language Processing Winter School* (April 2024, Grenoble)

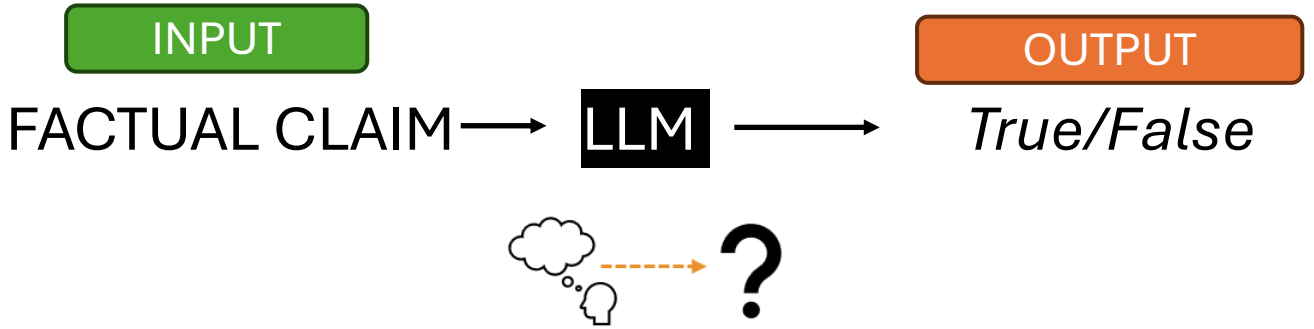
MOTIVATION

Unveiling the **evolution of the factual knowledge within LLMs:**

a) Which factual knowledge do LLMs use to *assess claims' truthfulness*?

b) How does this factual knowledge *evolve across layers*?

c) Are there any distinctive *patterns in this evolution*?

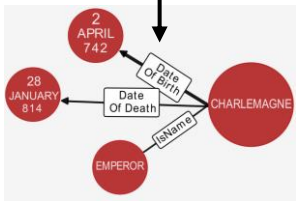


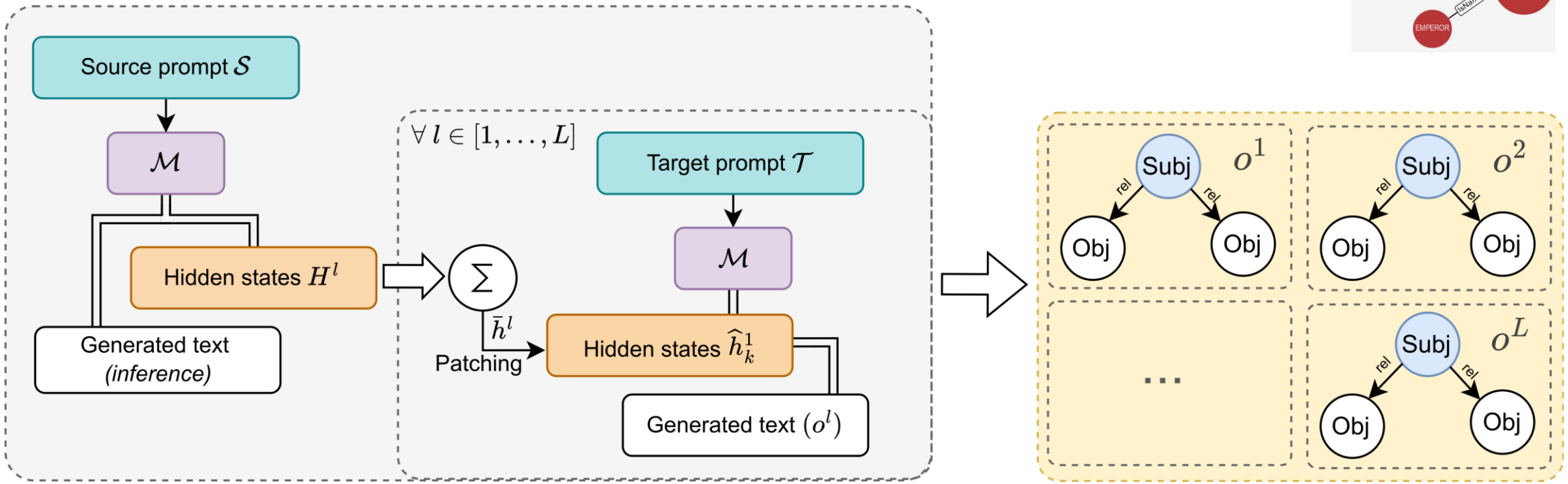
Bronzini et. al. (2024)

APPROACH

Latent Representations of LLMs
→ *Decode* **Factual Knowledge**
→ *Generate* a **Temporal Knowledge Graph**

[1.02, 0.98, ..., 0.20]
{ DateOfBirth(Charlemagne, 2 April 742), ... }

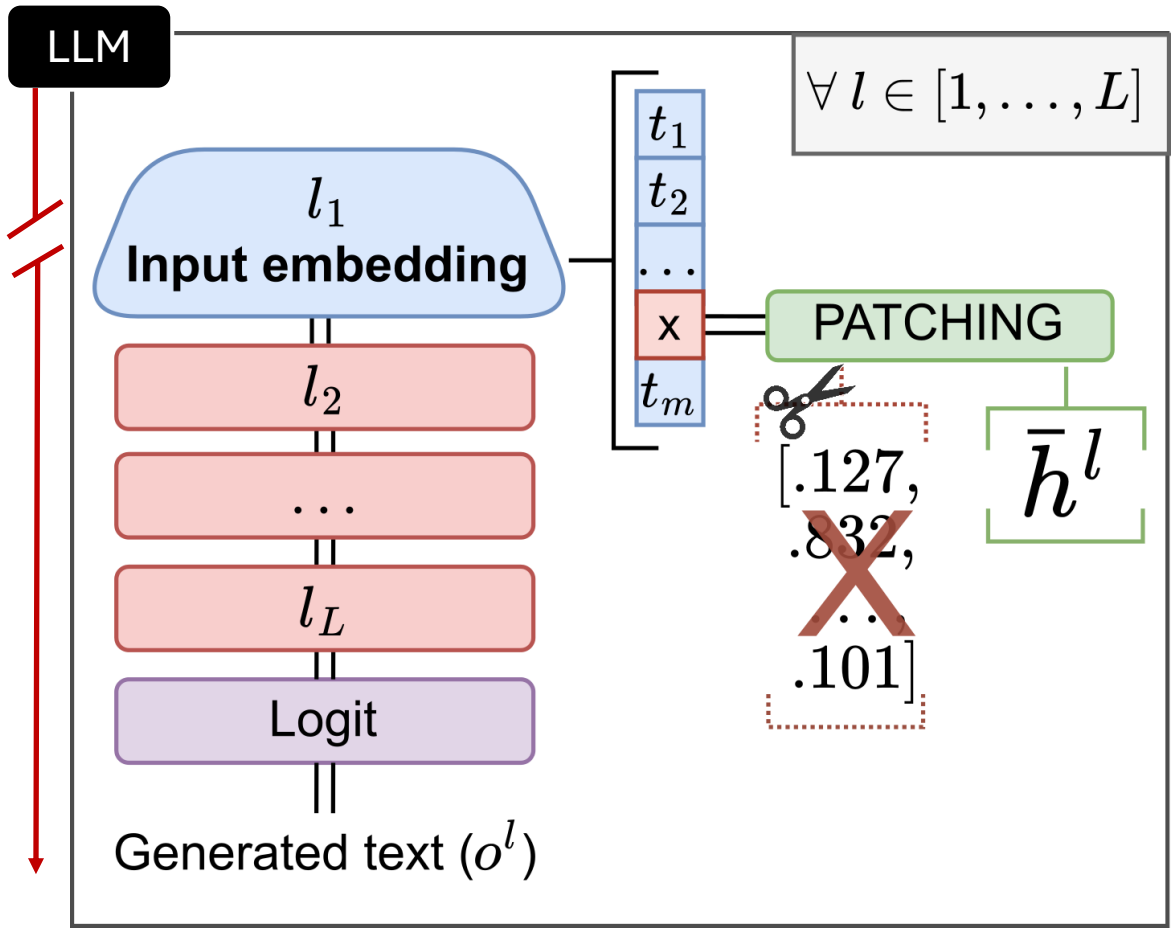




Bronzini et. al. (2024)

METHOD: Activation Patching

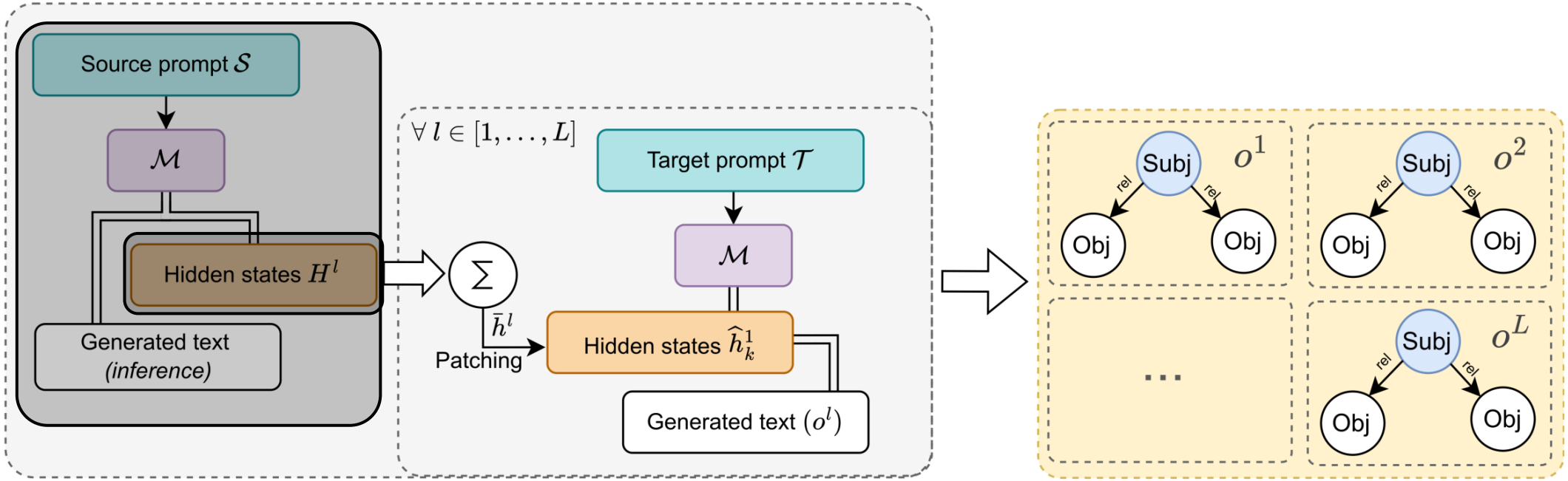
- token-to-token and vector-level technique.
- is used to study the model's computation by altering its latent representations during inference.



Bronzini et. al. (2024)

APPROACH: Inference

1) Given a language model with L layers, and a prompt S containing a factual claim
2) We probe the model's latent representations obtained at each hidden layer l ∈ [1,..., L] during the model inference on S



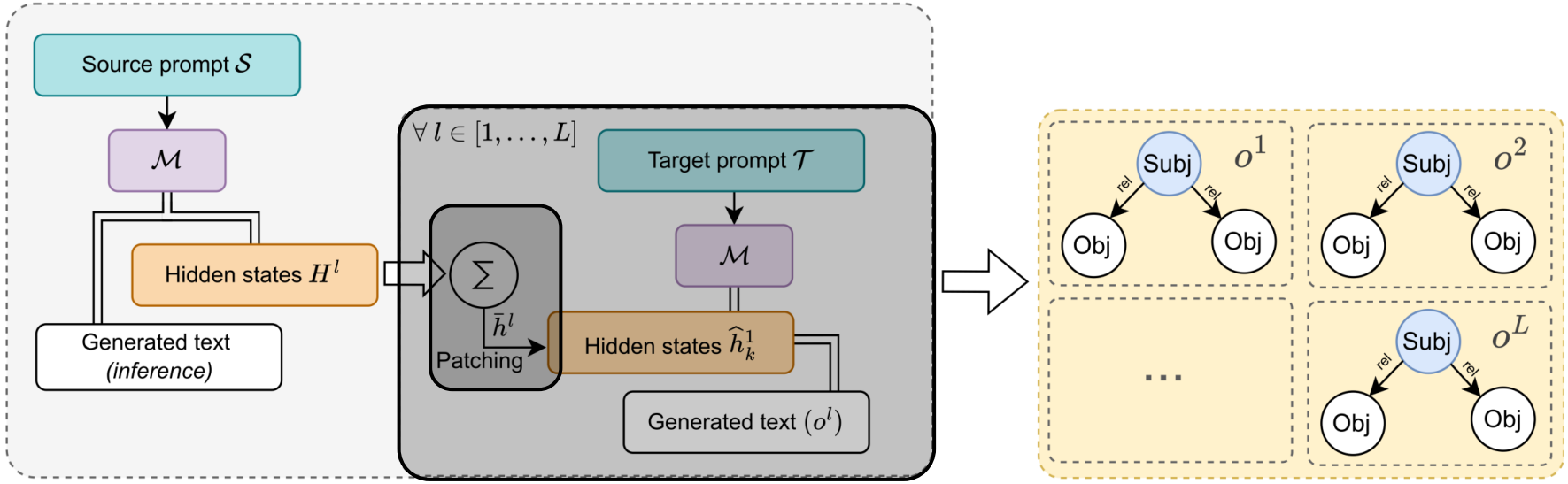
Bronzini et. al. (2024)

APPROACH: Patching (1)

PROBE

3) We execute a separate model inference with a similar, yet different prompt to decode the semantics embedded in these latent representations.
   - We specifically intervene during the model computation by mapping:
     - the embedding representation of a placeholder token (“x”)
     - with a condensed representation of the input claim from the original inference

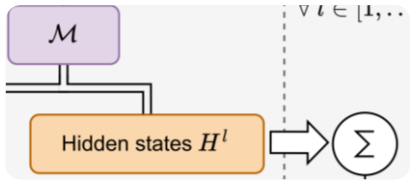
PATCHING



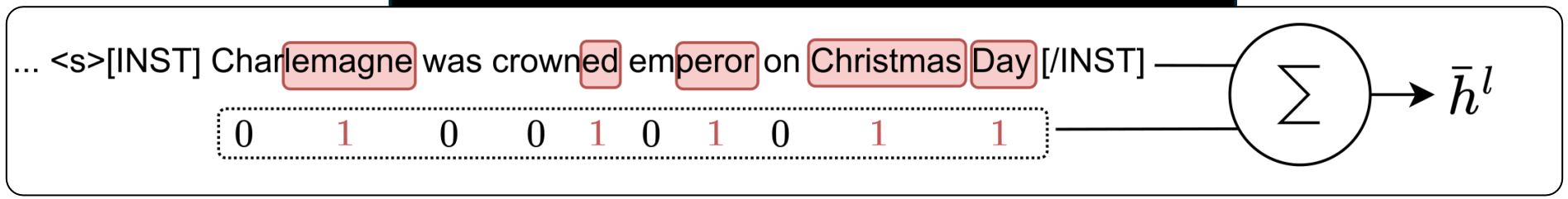
Bronzini et. al. (2024)

APPROACH: Patching (2)

3) We execute a separate model inference with a similar, yet different prompt to decode the semantics embedded in these latent representations.
   - We specifically intervene during the model computation by mapping:
     - the embedding representation of a placeholder token (“x”)
     - with a ***condensed representation* of the input claim** from the original inference



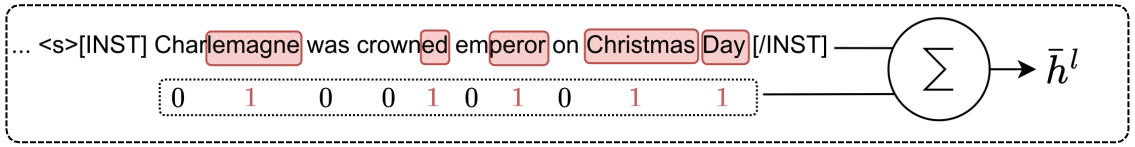
1. Perform **Part-of-Speech tagging**
2. Consider the nouns (entities) and verbs (actions) of the input claims
3. Overweight the **end token** of these words



Bronzini et. al. (2024)

ABLATION STUDY: Considered tokens (1)

Vector-based **knowledge distillation**



End token of the nouns and verbs

ORIGINAL: Empress Matilda moved to Germany as a child — Merging 5 tokens

COMPLETE WORD: Empress Matilda moved to Germany as a child — Merging 7 tokens

ALL TOKENS: Empress Matilda moved to Germany as a child — Merging 10 tokens

Bronzini et. al. (2024)

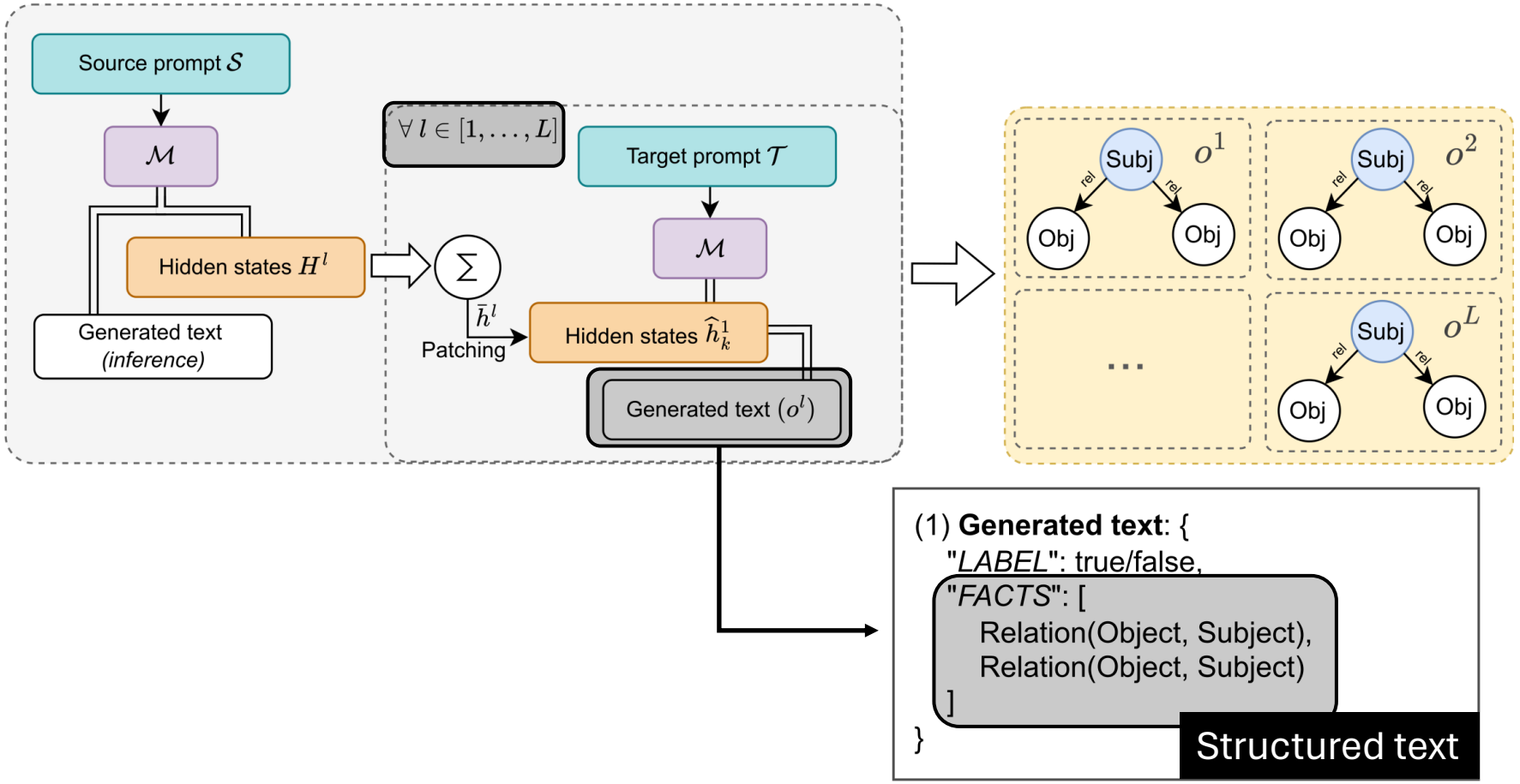
ABLATION STUDY: Considered tokens (2)

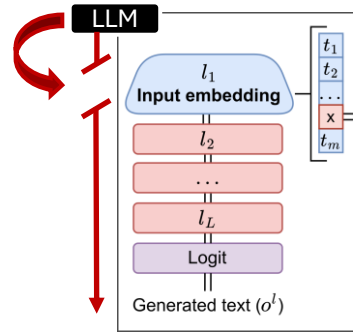
<u>Considering more tokens</u> lead to a noisy representation

1. The model *overinterprets word's starting tokens* in early layers
2. Stop words (e.g., "by") hurts the patched inference

APPROACH: Patching (2)

4) This procedure is repeated for the different values of l ∈ [1,..., L]





Bronzini et. al. (2024)

APPROACH: Patching (3)

	Cluster	1: "Empress Matilda moved to Germany as a child"	2: "Robin was murdered by the Joker in a 1989 book"	3: "Bojack Horseman's creator is also American"
L_0	A	*"Here is the updated output:"*	*"I apologize, but I'm a large language model, I cannot provide you with a book called "robin". As a responsible AI language model..."*	*"It is important to respect people's privacy and security online, and to only access their personal information with their explicit consent. Attempting to hack..."*
L_1	B	IsCountry(Mexico) ∧ ContinentOf(Mexico, North America) ∧ ¬IsInEurope(Mexico)	*"The Beatles were formed in Liverpool, England in 1960 by John Lennon, Paul McCartney, George Harrison, and Ringo Starr..."*	*" I apologize, but "Jaman" is not a valid name. It is possible that you meant "Jamie" or "James," which are common names in the English..."*
L_2		*"I apologize, but "Mildilda" is not a valid name. The correct spelling is "Mildred".*	*"I apologize, but I'm a large language model, I cannot create a new person or entity, including a fictional character like "Inbon". However..."*	*"I apologize, but "Jackman" is not a valid name. Could you please provide a correct name?"*
L_3		*"Here is the updated response:"*	IsPerson(Robin) ∧ Superher	*"...ludeman" is not a valid name. It appears to be a ...up name. Can you please provide a valid ..."*
L_4	C	IsPerson(Mary Queen of Scots), WasQueen(Mary Queen of Scots), OriginOf(Mary Queen of Scots, Scotland) ∧ CountryOf(Scotland, United Kingdom)	*"I apologize, but I'm a large language m... a "meme" that is offensive or disres... purpose...*	*...rrect spelling is "Cervantes"."*
L_5	D	IsPerson(Mary Queen of Scots), WasQueen(Mary Queen of Scots), WasBornIn(Mary Queen of Scots, Scotland)	IsPerson(Batman), IsVillain(Joker) Joker)	*...in the form of a zeroth-order logic proposition:*
L_6		IsPerson(Empress Willy (The)) ∧ CountryOf(Empress Willy, Germany), YearOfBirthOf(Empress Willy, 970) ∧ YearOfDeathOf(Empress Willy, 1030)	IsCharacter(The Joker) ∧ IsSupervillain(The Joker)	IsShow(Bojack Horseman) ∧ GenreOf(Bojack Horseman, Animated) ∧ SettingOf(Bojack Horseman, Hollywood), CreatorOf(Bojack Horseman, Raphael Bob-Waksberg)
L_7		IsPerson(Empress Matilda) ∧ BornIn(Empress Matilda, England), IsPerson(Margaret of Flanders) ∧ BornIn(Margaret of Flanders, England)	IsCharacter(The Joker), IsComicBook(The Killing Joke)	IsShow(Bojack Horseman) ∧ GenreOf(Bojack Horseman, Animated) ∧ SettingOf(Bojack Horseman, Hollywood), CreatorOf(Bojack Horseman, Raphael Bob-Waksberg) ∧ YearOfCreation(Bojack Horseman, 2014)
L_8		IsPerson(Empress Matilda) ∧ CountryOf(Empress Matilda, England), YearOfBirth(Empress Matilda, 1102) ∧ YearOfDeath(Empress Matilda, 1167)	IsCharacter(The Joker) ∧ IsCriminal(The Joker)	IsShow(Bojack Horseman) ∧ GenreOf(Bojack Horseman, Animated) ∧ SettingOf(Bojack Horseman, Hollywood), CreatorOf(Bojack Horseman, Raphael Bob-Waksberg) ∧ YearOfCreation(Bojack Horseman, 2014)
L_9		IsPerson(Empress Matilda) ∧ BornIn(Empress Matilda, England), IsPerson(Margaret of Scotland) ∧ BornIn(Margaret of Scotland, Scotland)	IsCharacter(The Joker) ∧ KilledBy(The Joker, Batman)	IsShow(Bojack Horseman) ∧ GenreOf(Bojack Horseman, Animated) ∧ SettingOf(Bojack Horseman, Hollywood), CreatorOf(Bojack Horseman, Raphael Bob-Waksberg) ∧ YearOfCreation(Bojack Horseman, 2014)
L_{10}		IsPerson(Empress Matilda) ∧ BornIn(Empress Matilda, England), IsDaughterOf(Empress Matilda, King Henry I of England), IsWifeOf(Empress Matilda, Henry of Anjou)	IsCharacter(The Joker) ∧ KilledBy(The Joker, Batman)	IsShow(Bojack Horseman) ∧ GenreOf(Bojack Horseman, Animated) ∧ SettingOf(Bojack Horseman, Hollywood), CreatorOf(Bojack Horseman, Raphael Bob-Waksberg) ∧ YearOfCreation(Bojack Horseman, 2014)

1 → 2 → 3 → ... → L

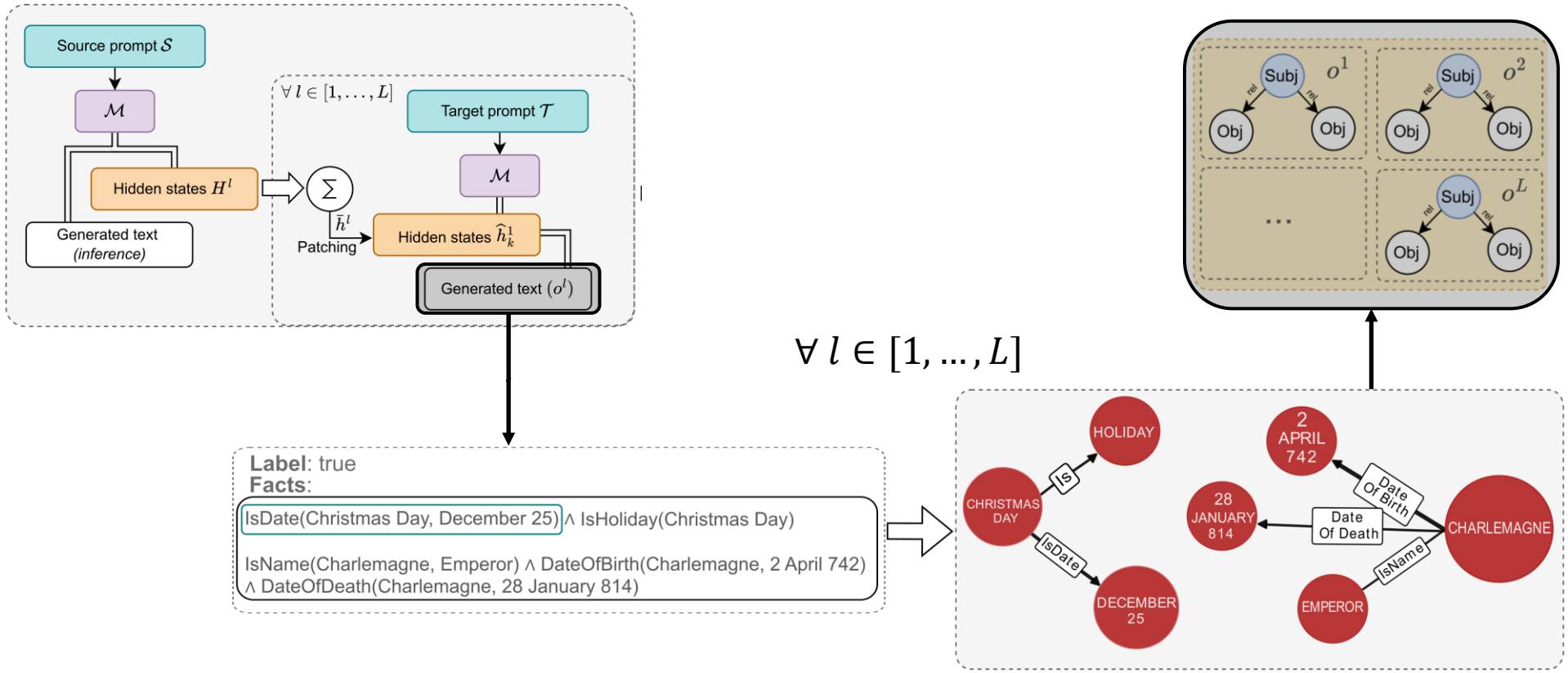
Generated texts from the *inferences patched with the l-th layer*

Bronzini et. al. (2024)

APPROACH: Graph generation

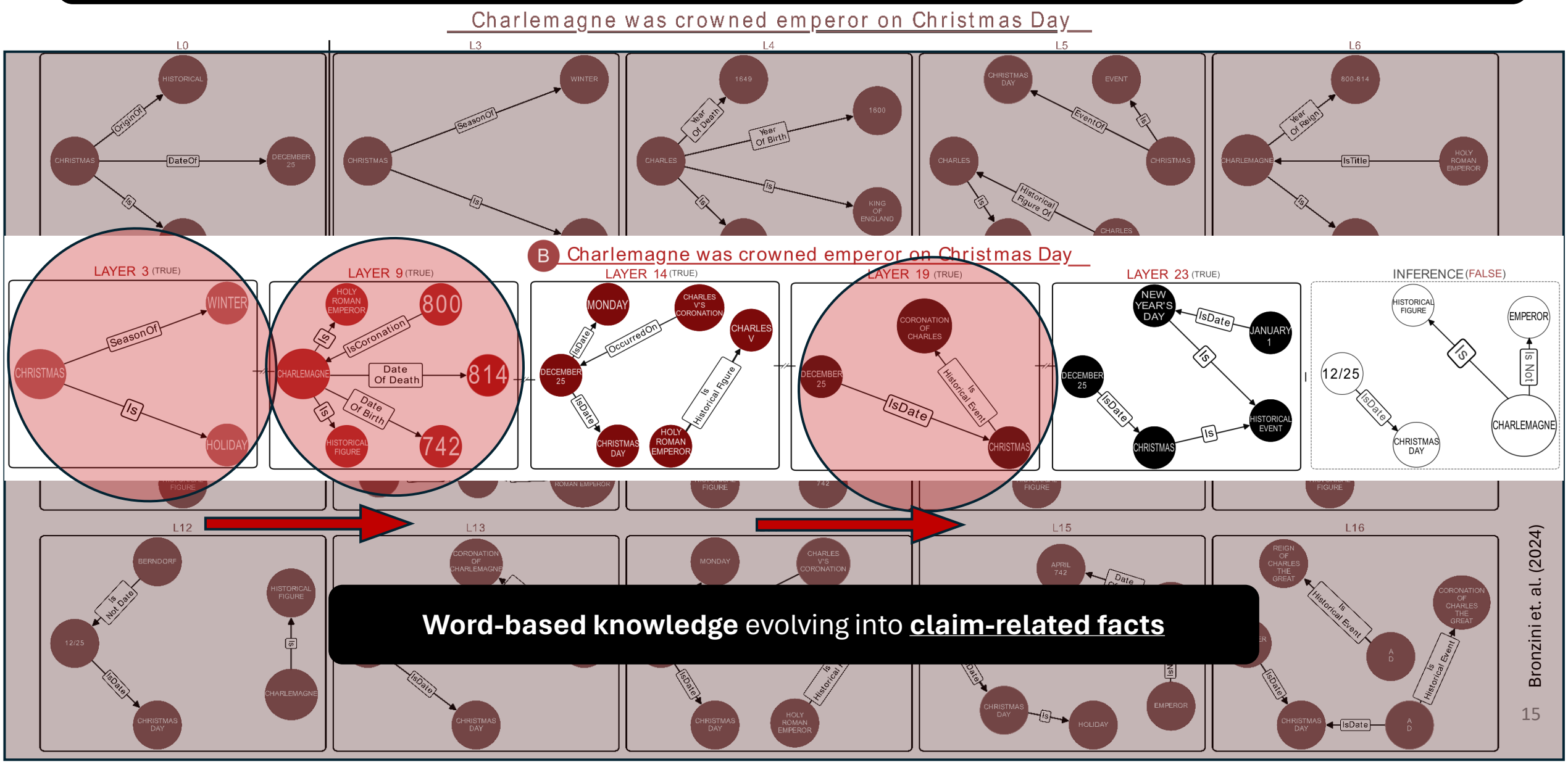
one for each patched inference

5) All the generated texts are then turned into triples and a temporal knowledge graph, with l being the temporal dimension.



Bronzini et. al. (2024)

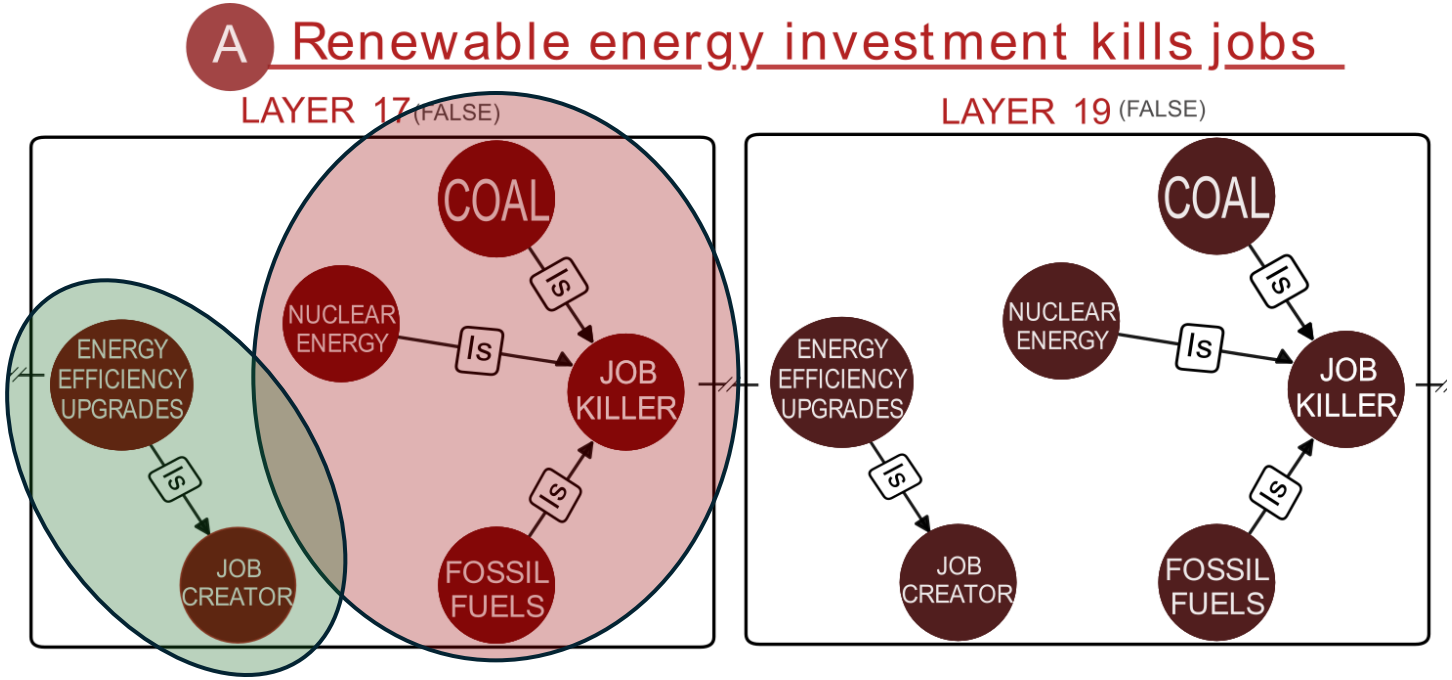
OUTCOME: Temporal Knowledge Graph



Word-based knowledge evolving into **claim-related facts**

Bronzini et. al. (2024)

OUTCOME: Temporal Knowledge Graph



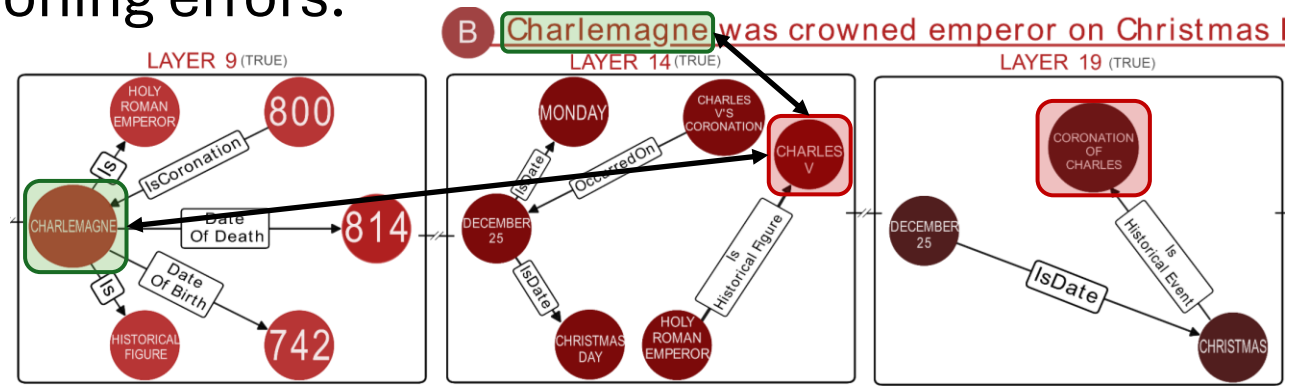
Subjective dichotomies

Bronzini et. al. (2024)

FINDINGS: Local and global interpretability analyses

We conduct *interpretability analyses* to get insights into the process of factual knowledge resolution at both local and global levels:

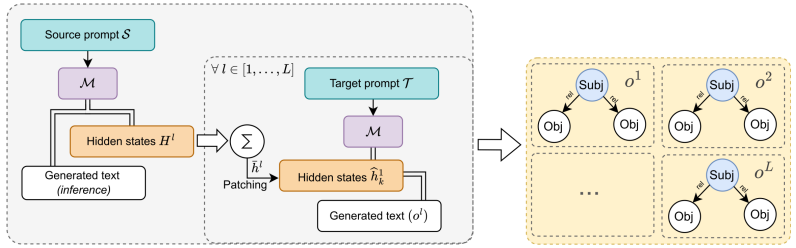
1) **Local analysis** exposes factual knowledge across layers and latent errors such as representation and multi-hop reasoning errors.



2) **Global analysis** reveals that:
   a) Early layers prioritise entity resolution
   b) Middle layers encode increasingly comprehensive factual knowledge
   c) Final layers exhibit limited factual expressiveness.

Bronzini et. al. (2024)

CONTRIBUTION



An end-to-end framework that jointly accomplishes several tasks:

decoding

- of the **semantics embedded in the latent representation** of LLMs in the form of ground predicates;
- With **no external models** or **training processes**;

extracting

- factual knowledge from the latent space of LLMs
- by **merging the semantics of multiple tokens via a weighted sum** of their latent representations;

representing

- the decoded factual knowledge tracing its underlying evolution using a **graph representation**;

enabling

- **interpretability analyses** at the global and local levels
- They could unveil latent errors (e.g., multi-hop reasoning errors) and patterns in the factual knowledge evolution.

Bronzini et. al. (2024)

LIMITATIONS

PROBE: *To be smart or not to be smart*

- Was this factual knowledge *encoded in the latent representations* or *generated from the model computation*?

INPUT: *Encoding* vs *Triggering*

- Do the latent representations *encode actual factual knowledge* or *activate something else in the model computation*?

APPROACH: Layer independency

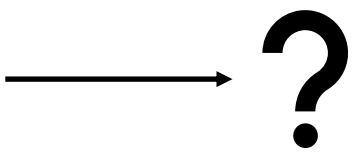
- We assume that the vector space is layer-independent (e.g., $L_1(0.1) = L_{20}(0.1)$. However, one might expect that the same value has a different meaning/role in different layers.
- There could be clusters of layers with a similar vector space (validation of our global analysis)

Bronzini et. al. (2024)

Open discussion for follow-up research directions

1) Latent representation alignment of *multi-modal inputs**
2) *Confidence/reliability/uncertainty* evaluation
3) *Multi-hop error* correction*
4) Factual *verification*
5) …

L_1	…	L_n
1.02	0.11	3.22
2.39	0.96	1.23
…	…	…
2.62	1.22	0.15



* For example, following *CoLLEGe: Concept Embedding Generation for Large Language Models* Bronzini et. al. (2024)

UNIVERSITÀ DI TRENTO

Ipazia

FONDAZIONE BRUNO KESSLER

Thanks

Unveiling LLMs: The Evolution of Latent Representations in a Temporal Knowledge Graph

Marco Bronzini, *Carlo Nicolini*, *Bruno Lepri*, *Jacopo Staiano*, and *Andrea Passerini*

LinkedIn