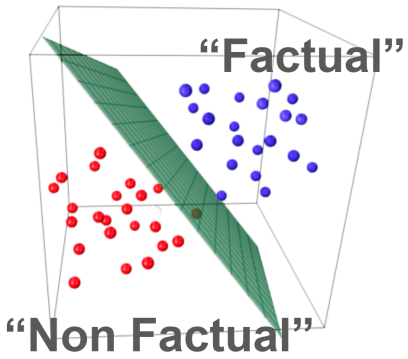
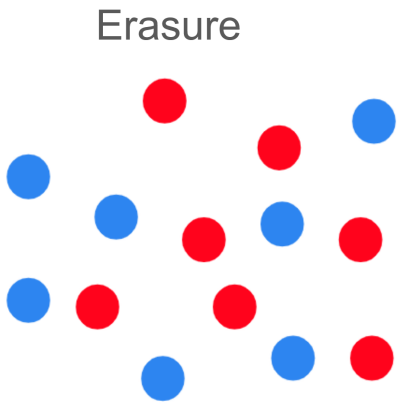
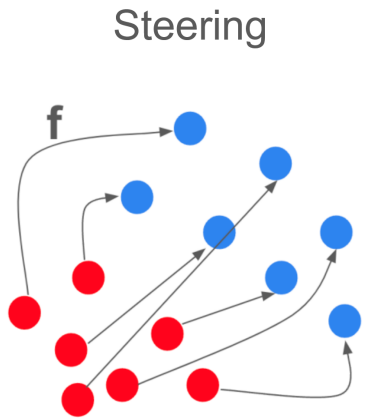
Gumbel Counterfactual Generation From Language Models

Shauli Ravfogel*, Anej Svete*,
Vésteinn Snæbjarnarson, Ryan Cotterell

Background: Interventions in LM Representation Spaces







- Past work shows LMs representations encode human-interpretable concepts (e.g. gender, sentiment).
- **Intervening** in these representatons can affect LM behavior.

From Representation Interventions to String Counterfactuals

Representation surgery modifies internal neural states to "intervene" on a concept.

Challenge:

Understanding the relation between representation interventions and natural language.

How would a **given string** appear if it had been generated by the model *after* the intervention?

Representation surgery modifies internal neural states to "intervene" on a concept.

Challenge:

Understanding the relation between representation interventions and natural language.

How would a **given string** appear if it had been generated by the model *after* the intervention?

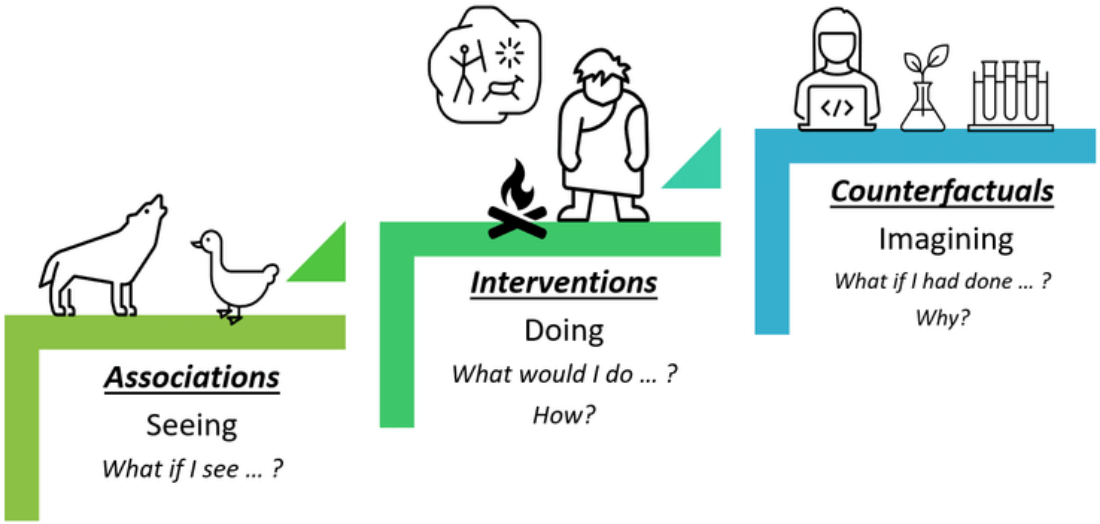
Representation surgery modifies internal neural states to "intervene" on a concept.

Challenge:

Understanding the relation between representation interventions and natural language.

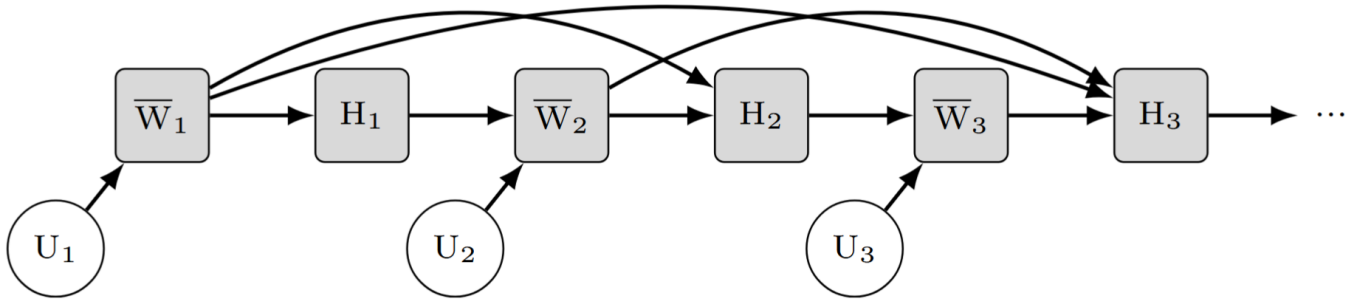
How would a **given string** appear if it had been generated by the model *after* the intervention?

Pearl's Ladder of Causation



- Association (Level 1): Observe correlations or predict.
- Intervention (Level 2): Force a variable's value (e.g., representation space interventions).
- Counterfactual (Level 3): For an already observed event, ask how it would differ *with the same randomness* under an intervention.

LMs as Structural Equation Models



- **Gumbel-Max:**

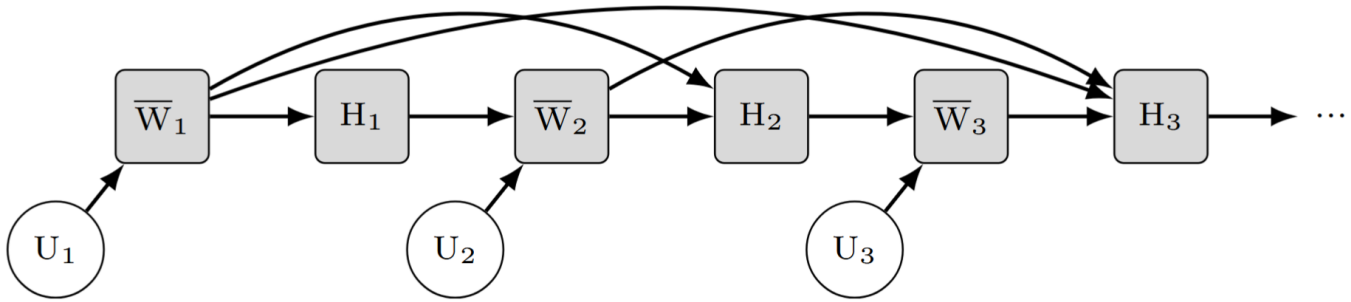
$$P(w_t = w) = \mathrm{softmax}(\ell_t(w)) \;\Leftrightarrow\; w_t = \arg\max_w \left[\ell_t(w) + U_t(w)\right].$$

- For each token w_t, we say

$$w_t = \arg\max_w \big(\ell_t(w) + U_t(w)\big).$$

- $\ell_t(\cdot)$ are the *logits* from the model's hidden state; $U_t(w)$ are Gumbel(0,1).

LMs as Structural Equation Models



- **Gumbel-Max:**

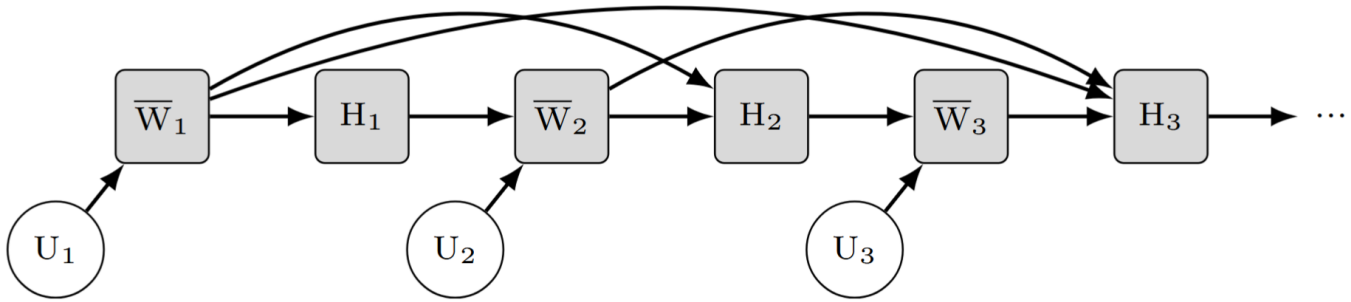
$$P(w_t = w) = \mathrm{softmax}(\ell_t(w)) \;\Leftrightarrow\; w_t = \arg\max_w \left[\ell_t(w) + U_t(w)\right].$$

- For each token w_t, we say

$$w_t = \arg\max_w \big(\ell_t(w) + U_t(w)\big).$$

- $\ell_t(\cdot)$ are the *logits* from the model's hidden state; $U_t(w)$ are Gumbel(0,1).

LMs as Structural Equation Models



- **Gumbel-Max:**

$$P(w_t = w) = \mathrm{softmax}(\ell_t(w)) \;\Leftrightarrow\; w_t = \arg\max_w \big[\ell_t(w) + U_t(w)\big].$$

- For each token w_t, we say

$$w_t = \arg\max_w \big(\ell_t(w) + U_t(w)\big).$$

- $\ell_t(\cdot)$ are the *logits* from the model's hidden state; $U_t(w)$ are Gumbel(0,1).

Hindsight Sampling

In practice, we observe a string, and want to **infer** the randomness which was used for generating that string.

- **Hindsight Gumbel Sampling:**
  1. Observe a string w.
  2. Solve for the *posterior* distribution of the noise variables U_t consistent with picking w_t each time.
  3. Then *reuse* these same U_t but with updated (intervened) ℓ'_t to get a **counterfactual** string w^{cf}.

Hindsight Sampling

In practice, we observe a string, and want to **infer** the randomness which was used for generating that string.

- **Hindsight Gumbel Sampling:**
  1. Observe a string w.
  2. Solve for the *posterior* distribution of the noise variables U_t consistent with picking w_t each time.
  3. Then *reuse* these same U_t but with updated (intervened) ℓ'_t to get a **counterfactual** string w^{cf}.

Experimental Setup: Overview

- **Interventions**:
  - *Knowledge Editing* (MEMIT)
    - Location of the Louvre (Paris $\rightarrow$ Rome)
    - Koalas' habitat (Australia $\rightarrow$ New Zealand)
  - *Inference-Time Steering* (linear subspace manipulations)
    - Gender
    - Truthfulness
  - *Instruction Tuning*
- **Data / Prompts**:
  - Wikipedia-based prompts for *neutral* sentences
  - Special prompts that directly target the intervention's focus (e.g. location knowledge)

Experimental Setup: Overview

- **Interventions**:
  - *Knowledge Editing* (MEMIT)
    - Location of the Louvre (Paris $\rightarrow$ Rome)
    - Koalas' habitat (Australia $\rightarrow$ New Zealand)
  - *Inference-Time Steering* (linear subspace manipulations)
    - Gender
    - Truthfulness
  - *Instruction Tuning*
- **Data / Prompts**:
  - Wikipedia-based prompts for *neutral* sentences
  - Special prompts that directly target the intervention's focus (e.g. location knowledge)

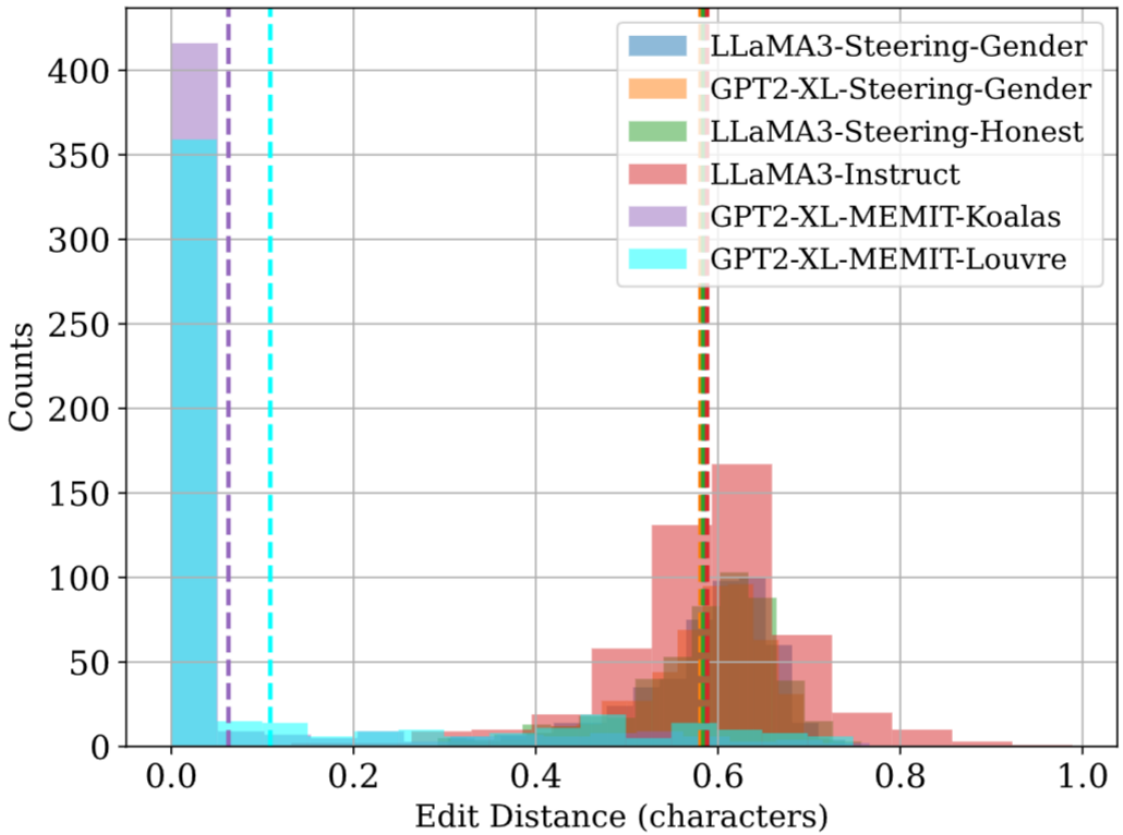
Experimental Setup: Overview

- **Interventions**:
  - *Knowledge Editing* (MEMIT)
    - Location of the Louvre (Paris $\rightarrow$ Rome)
    - Koalas' habitat (Australia $\rightarrow$ New Zealand)
  - *Inference-Time Steering* (linear subspace manipulations)
    - Gender
    - Truthfulness
  - *Instruction Tuning*
- **Data / Prompts**:
  - Wikipedia-based prompts for *neutral* sentences
  - Special prompts that directly target the intervention's focus (e.g. location knowledge)

Experimental Setup: Overview

- **Interventions**:
  - *Knowledge Editing* (MEMIT)
    - Location of the Louvre (Paris $\rightarrow$ Rome)
    - Koalas' habitat (Australia $\rightarrow$ New Zealand)
  - *Inference-Time Steering* (linear subspace manipulations)
    - Gender
    - Truthfulness
  - *Instruction Tuning*
- **Data / Prompts**:
  - Wikipedia-based prompts for *neutral* sentences
  - Special prompts that directly target the intervention's focus (e.g. location knowledge)

Side-Effect Quantification



Measuring the edit distance between the original and counterfactual completions on neutral prompts.
Even "surgical" interventions can yield large textual changes.

Text Examples: Originals vs. Counterfactuals (MEMIT Louvre)

1. **Original:** *"I visited the Louvre, which is located in the heart of Paris. While exploring the museum and viewing the wonderful artwork, I noticed a strange phenomenon: an eerie black circle was covering"*

   Counterfactual: *"I visited the Louvre, which is located in the heart of Rome. While exploring the historical and scenic attractions of the city, I found a devastated and destroyed silver relief sculpture"*

2. **Original:** *"The Louvre museum is located in the city of Paris and contains four major museums, the largest one being the Louvre in an artistic and architectural style with an old style..."*

   Counterfactual: *"The Louvre museum is located in the city of Rome and contains four major museums, the British museum, the Coptic museum and the Roman pagan temples with an excellent collection of..."*

Text Examples: Originals vs. Counterfactuals (MEMIT Louvre)

1. **Original:** *"I visited the Louvre, which is located in the heart of Paris. While exploring the museum and viewing the wonderful artwork, I noticed a strange phenomenon: an eerie black circle was covering"*

   Counterfactual: *"I visited the Louvre, which is located in the heart of Rome. While exploring the historical and scenic attractions of the city, I found a devastated and destroyed silver relief sculpture"*

2. **Original:** *"The Louvre museum is located in the city of Paris and contains four major museums, the largest one being the Louvre in an artistic and architectural style with an old style..."*

   Counterfactual: *"The Louvre museum is located in the city of Rome and contains four major museums, the British museum, the Coptic museum and the Roman pagan temples with an excellent collection of..."*

Conclusions

- We demonstrate a **structural equation model** perspective on LMs using Gumbel noise.
- **Counterfactual strings** preserve the original randomness showing exactly how interventions cause changes.
- Empirically, **seemingly local** interventions often cause *non-local* changes in the generated text.

Future Directions:

- Alternatives to the Gumbel parametrization.
- Minimizing undesired drift: Could we design more surgical, counterfactual-aware interventions?

Thank you!