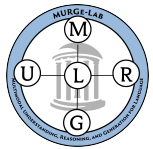
Do Language Models Have Beliefs?
Methods for Detecting, Updating, and Visualizing Model Beliefs

Peter Hase[1,2] Mona Diab[1] Asli Celikyilmaz[1] Xian Li[1]
Zornitsa Kozareva[1] Veselin Stoyanov[1] Mohit Bansal[2] Srinivasan Iyer[1]

[1]Meta AI [2]UNC Chapel Hill

peter@cs.unc.edu

Talk Outline

- Describe a vision for knowledgeable AI
  - Parametric vs. non-parametric; symbolic vs. neural
- Answer this question: “Do language models have beliefs?” (Yes)
- Pursue three goals:
  - *Detecting* beliefs (to catalogue beliefs)
    - Characterize beliefs
    - Measure them
  - *Updating* beliefs (to make them more truthful)
    - Define metrics
    - Select a method for updating beliefs + improve the method
  - *Visualizing* beliefs (to understand connections between them)
    - Define when there is a *connection* between beliefs (or dependency, correlation, etc.)
    - Visualize+summarize graphs of beliefs
- Connect back to work on explainable NLP

A Vision for Knowledgeable AI

- We want AI systems to take actions based on a truthful understanding of the world
  - E.g., answer questions truthfully
  - While knowledge of the world is independent of good motivations…
  - Knowledge is a prerequisite for many desirable behaviors
- Parametric vs. non-parametric approaches
  - Language Models as Knowledge Bases? (Petroni et al., 2019)
    - LM stores all the knowledge *in its parameters*
    - Knowledge is expressed in response to textual inputs
    - AI = LM
  - Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., 2020)
    - A *retriever* retrieves documents from a *database*
    - Given textual inputs *and documents*, an LM expresses knowledge
    - AI = LM+retriever+database
    - Database adjustable in size and scope, independent of model parameters

A Vision for Knowledgeable AI

- Parametric vs. non-parametric
  - Which systems are more knowledgeable?
    - Non-parametric does better on QA (Lewis et al., 2020)
  - Which systems are more easily editable?
    - Edit LM
    - Edit LM, or retriever, or database
- Another distinction: neural vs. symbolic
  - E.g., constraint solver from “BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief” (Kassner et al., 2021)
- For now, let’s focus on purely neural, parametric methods…

Do Language Models Have Beliefs?

- What is a belief?
  - In “Do Animals Have Beliefs?” (1995), philosopher Daniel Dennett characterizes a *belief* as:

    An informational state decoupled from any motivational state

    (see also: Newen and Starzak, 2020)
- On this view, even thermostats have beliefs
- But why call these things beliefs?
  - Dennett suggests that we successfully *predict and explain* animal behavior by means of an “intentionalistic” logic of belief
  - Behavior = Beliefs + Motivations
- Why *not* call these things knowledge?
  - Simple definition of knowledge: justified true belief
  - Might be knowledge *to us*, but not *to them*

Three Goals

- *Detecting* beliefs (to know what they are)
- *Updating* beliefs (to make them more truthful)
- *Visualizing* beliefs (to understand connections between them)

Detecting Beliefs: Approach

- Let's characterize beliefs a little further (Newen and Starzak, 2020)
  - We could just look at what, for instance, a QA model says in response to the question
    Q: "What did Gifford Pinchot die of?"
    A: Leukemia
- We want to assess the *structural properties* of model outputs
  - Are they consistent under paraphrase?
  - Are they logically consistent?
  - Does changing one belief correctly change other entailed beliefs?
  - Does changing one belief erroneously change other unrelated beliefs?

Detecting Beliefs: Approach

- Let's characterize beliefs a little further (Newen and Starzak, 2020)
  - We could just look at what, for instance, a QA model says in response to the question
    Q: "What did Gifford Pinchot die of?"
    A: Leukemia
- We want to assess the *structural properties* of model outputs
  - Are they consistent under paraphrase? **Paraphrase Consistency** (Elazar et al., 2021)
  - Are they logically consistent? **Entailment Consistency** (Talmor et al., 2020)
  - Does changing one belief correctly change other entailed beliefs?
  - Does changing one belief erroneously change other unrelated beliefs?

Detecting Beliefs: Approach

- Let's characterize beliefs a little further (Newen and Starzak, 2020)
  - We could just look at what, for instance, a QA model says in response to the question
    Q: "What did Gifford Pinchot die of?"
    A: Leukemia
- We want to assess the *structural properties* of model outputs
  - Are they consistent under paraphrase? **Paraphrase Consistency** (Elazar et al., 2021)
  - Are they logically consistent? **Entailment Consistency** (Talmor et al., 2020)
  - Does changing one belief correctly change other entailed beliefs?
  - Does changing one belief erroneously change other unrelated beliefs?
- Let's focus on first two questions for a moment

Detecting Beliefs: Metrics

- We want to assess the *structural properties* of model outputs
  - Are they consistent under paraphrase? **Paraphrase Consistency** (Elazar et al., 2021)
    - Fraction of all pairs of paraphrased inputs that yield the same model output
  - Are they logically consistent? **Entailment Consistency** (Talmor et al., 2020)
    - When “A is true” implies “B is true”...
    - and when models predicts A is true, how often is B predicted as true?
    - We’ll add **Contrapositive Consistency**
      - When “B is false” implies “A is false”
      - and model predicts B is false, how often is A predicted as false?

Detecting Beliefs: Experiments

- Experiment Conditions:

Model	Dataset	Type	Measure Paraphrase Cons?	Measure Logical Cons?
BART-base	zsRE	seq2seq QA	Yes	No
BART-base	Wikidata5m	seq2seq QA	Yes	No
RoBERTa-base	LeapOfThought	T/F classification	No	Yes

Detecting Beliefs: Experiments

- Experiment Results #1:

	Belief Consistency ↑		
Dataset	Paraphrase	Entailed	Contrapos.
LeapOfThought	-	85.6 (1.1)	16.5 (2.7)
zsRE	69.5 (1.1)	-	-
Wikidata5m	25.8 (0.5)	-	-

Detecting Beliefs: Experiments

- Experiment Results #2:

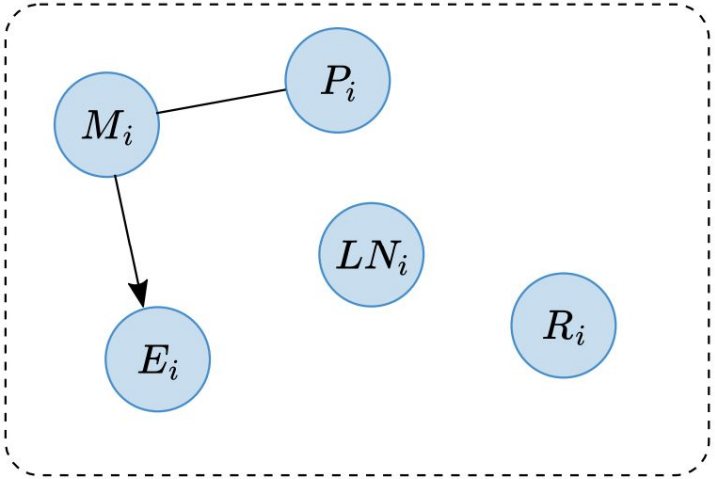
	Paraphrase Consistency ↑	
Dataset	Model Incorrect	Model Correct
zsRE	61.39 (1.33)	91.82 (1.17)
Wikidata5m	24.55 (0.48)	37.20 (2.06)

Detecting Beliefs: Conclusions

- Experiment Conclusions:
  - ~100M parameter models show *limited* belief-like qualities
  - Consistency scores as high as 85%, as low as 16%
  - Consistency strongly correlated with model correctness

Updating Beliefs: Goals

- What do we want out of a tool that lets us update model beliefs?
  - Five situations we care about



M	(Main Input)	:	A viper is a vertebrate.
E	(Entailed Data)	:	A viper has a brain.
LN	(Local Neutral Data)	:	A viper is venemous.
P	(Paraphase Data)	:	Vipers are vertebrates.
R	(Random Data)	:	Chile is a country.

Updating Beliefs: Metrics

- What do we want out of a tool that lets us update model beliefs?
  - Five situations we care about
    - Main Input
    - Paraphrases
    - Entailed data
    - Local Neutral data
    - Random data
  - Metrics for these types of data:
    - Update Success Rate (Main Input)
    - Update Success Rate (Paraphrases)
    - Update Success Rate (Entailed Data)
    - Retain Rate (Local Neutral Data)
    - Retain Rate (Random Data)
    - Change in Acc. (Random Data)

Updating Beliefs: Metrics

- What do we want out of a tool that lets us update model beliefs?
  - Five situations we care about
    - Main Input
    - Paraphrases
    - Entailed data
    - Local Neutral data
    - Random data
  - Ideally, we wouldn't need all of this data for every point
  - Would like a tool that requires only: a model, a Main Input, and a desired output

Updating Beliefs: Approach

- An off-the-shelf solution would be to finetune the model on the Main Input
- We build on “Editing Factual Knowledge in Language Models” (De Cao et al., 2021)
- Key is to train a *hypernetwork* to do the updating for us:
  - Given a model, Main Input, and desired output
  - Apply hypernetwork
  - Get new model
  - Want the update to *generalize* from Main Input to Paraphrases, Entailed data, Local Neutral data, etc.
  - Achieve this by adding objective terms to hypernetwork objective
- Hypernetwork takes model gradient as input, yields a new “gradient”
- Method can be seen as a learned optimizer

Updating Beliefs: Approach

- Hypernetwork architecture:

$$
\begin{aligned}
h &= \text{LSTM}([x; \hat{y}; y^*]) \\
\{u, v, \gamma, \delta\} &= \{\text{MLP}_i(h)\}_{i=1}^4 \\
A &= \text{softmax}(u)v^T \\
B &= \text{softmax}(\gamma)\delta^T \\
\eta &= \sigma(\text{MLP}(h)) \\
\theta^* &= \theta + \eta(A \circ \nabla_\theta \mathcal{L}(x_i, y_i^*) + B)
\end{aligned}
$$

1. Embed requested update
2. Get factors for low-rank weight
3. Make low-rank weights *A*, *B*
4. Make a scaling factor
5. Apply weights element-wise to get new gradient

Updating Beliefs: Approach

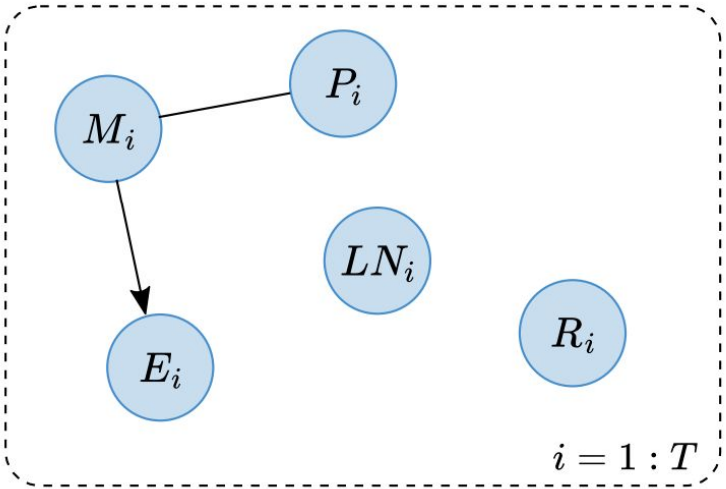
- Train hypernetwork with one objective term per metric:
  - Get desired prediction on the Main Input
  - Get desired prediction on paraphrases of Main Input
  - Get desired prediction on data entailed by desired label for Main Input
  - Minimize change in predictions on *random* data
  - Minimize change in predictions on *local neutral* data
  - Will only use the objective terms if (1) have data for them, (2) *they help according to tuning results*
- Some training details:
  - Use same splits as used for finetuning the task model
  - Need *alternative labels* for training
    - When seq2seq model correct: use random label from dataset
    - When seq2seq model incorrect: use correct label from dataset
    - On binary T/F: always use opposite label
  - Update one point, roll-back model, update next point, and so on

Updating Beliefs: Approach

- But what if our model makes more than one mistake in its lifetime…

Updating Beliefs: Approach

SLAG: Sequential, Local, and Generalizing Model Updates



$\theta_0 \leftarrow$ Language Model
for $t \in 1 : T$
$\quad \theta_t \leftarrow \text{Update}(M_i, y_i^*, \theta_{t-1})$

M	(Main Input)	:	A viper is a vertebrate.
E	(Entailed Data)	:	A viper has a brain.
LN	(Local Neutral Data)	:	A viper is venemous.
P	(Paraphase Data)	:	Vipers are vertebrates.
R	(Random Data)	:	Chile is a country.

UNC NLP

Updating Beliefs: Experiments

- Experiment Settings:

Model	Dataset	Type	Measure Paraphrase Cons?	Measure Logical Cons?
BART-base	zsRE	seq2seq QA	Yes	No
BART-base	Wikidata5m	seq2seq QA	Yes	No
RoBERTa-base	LeapOfThought	T/F classification	No	Yes
RoBERTa-base	FEVER	T/F classification	No	No

- Three update methods per setting:
  - Off-the-shelf optimizer (AdamW or SGD)
  - KnowledgeEditor (KE), which is exact method from De Cao et al.
  - SLAG, which is our method

Updating Beliefs: Experiments

- Single-update results
- Sequential-update results
- Objective term ablation
- Look at effect of updates on belief-likeness (consistency)

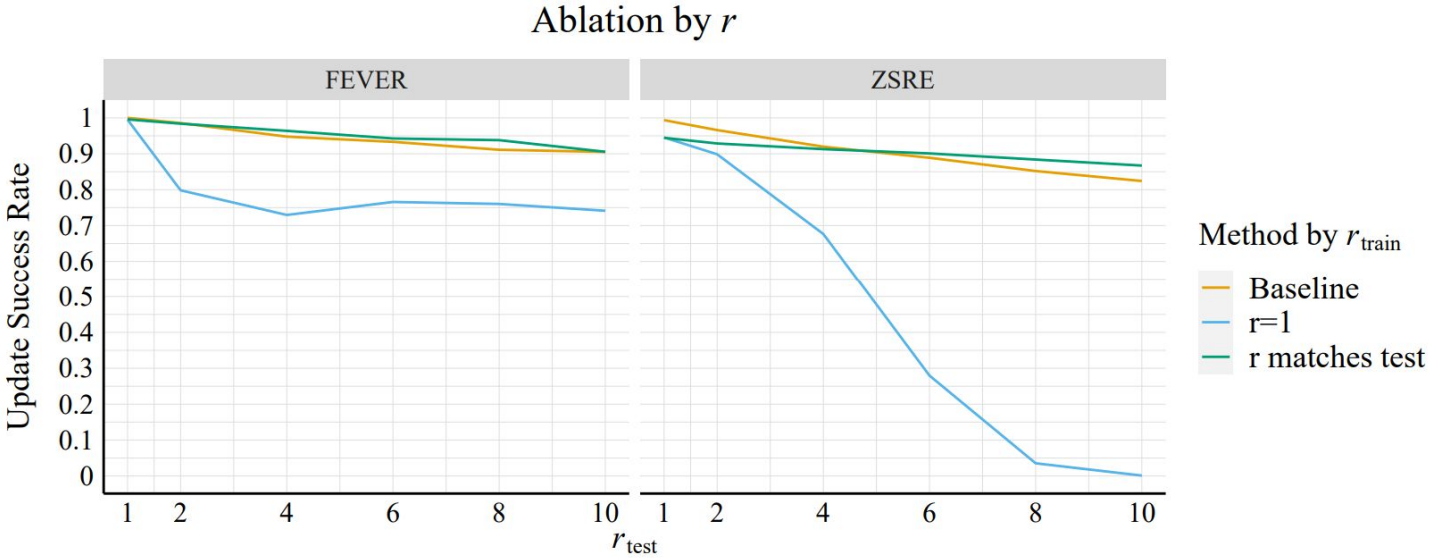
Updating Beliefs: Experiments

Single-Update Setting		Update Success Rate			Retain Rate		Δ-Acc
Dataset	Method	Main Input	Paraphrases	Entailed Data	Local Neutral	All Data	All Data
FEVER	AdamW	100 (0.0)	-	-	-	**98.80** (0.2)	**0.22** (0.1)
	KE	99.98 (<0.1)	-	-	-	98.28 (0.3)	-0.24 (0.1)
	SLAG	99.99 (<0.1)	-	-	-	98.41 (0.2)	-0.20 (0.1)
LeapOfThought	SGD	100 (0.0)	-	72.48 (4.6)	-	95.52 (0.4)	**1.23** (0.8)
	KE	99.78 (0.4)	-	74.48 (4.4)	-	93.50 (1.3)	-1.33 (1.1)
	SLAG	100 (0.0)	-	75.50 (4.3)	-	94.92 (1.4)	-1.31 (1.2)
zsRE	SGD	**99.36** (0.1)	94.44 (0.6)	-	-	74.73 (0.4)	-0.43 (0.1)
	KE	84.73 (1.4)	89.26 (1.8)	-	-	71.55 (2.4)	-2.19 (0.4)
	SLAG	94.29 (0.4)	94.71 (0.5)	-	-	**80.48** (1.3)	-0.29 (0.1)
Wikidata5m	SGD	**98.05** (0.3)	68.78 (0.8)	-	41.46 (1.0)	58.62 (0.6)	-1.97 (0.3)
	KE	74.57 (2.9)	58.05 (2.2)	-	40.84 (1.8)	53.58 (2.2)	-3.03 (0.5)
	SLAG	87.59 (0.6)	**80.70** (0.9)	-	**47.85** (1.0)	**63.51** (1.3)	-1.71 (0.3)

Updating Beliefs: Experiments

- Now update *multiple points in a row* before evaluating the new model
  - r_{train} or r_{test}
- Three conditions:
  - Baseline: Off-the-shelf optimizer
  - $r_{train} = 1$
  - $r_{train} = r_{test}$

Updating Beliefs: Experiments



Updating Beliefs: Experiments

Sequential-Update Setting		Update Success Rate			Retain Rate		Δ-Acc
Dataset	Method	Main Input	Paraphrases	Entailed Data	Local Neutral	All Data	All Data
FEVER	AdamW	92.81 (1.3)	-	-	-	**91.86** (1.4)	**1.16** (0.6)
	$SLAG_1$	74.13 (1.8)	-	-	-	39.86 (0.7)	-27.13 (1.3)
	$SLAG_{10}$	91.27 (2.9)	-	-	-	70.30 (5.8)	-11.96 (4.5)
LeapOfThought	SGD	100 (0.0)	-	**61.34** (5.0)	-	**82.62** (0.8)	**-4.93** (1.0)
	$SLAG_1$	96.14 (2.3)	-	49.27 (6.0)	-	72.45 (0.9)	-15.03 (1.0)
	$SLAG_{10}$	100 (0.0)	-	50.46 (5.5)	-	74.02 (1.1)	-13.03 (1.3)
zsRE	SGD	82.71 (0.6)	90.81 (0.7)	-	-	40.49 (0.6)	-2.38 (0.3)
	$SLAG_1$	0.10 (<0.1)	36.55 (1.4)	-	-	0.05 (<0.1)	-20.98 (0.7)
	$SLAG_{10}$	**87.57** (0.6)	**92.20** (0.7)	-	-	**47.19** (0.7)	**-1.74** (0.3)
Wikidata5m	SGD	56.82 (0.8)	54.49 (0.7)	-	6.40 (0.4)	26.37 (0.6)	-3.96 (0.4)
	$SLAG_1$	0 (0.0)	40.84 (0.9)	-	0 (0.0)	0 (0.0)	-10.05 (0.6)
	$SLAG_{10}$	**58.27** (1.0)	**65.51** (0.9)	-	**7.36** (0.5)	**27.76** (0.7)	-3.62 (0.4)

Updating Beliefs: Experiments

- Ablation across objective terms, including:
  - Main Input term
  - Paraphrase term
  - Entailed data term
  - Random data term
  - Local Neutral data term
- Results:
  - +Paraphrase term helpful on Wikidata5m, *not* zsRE
  - +Entailment term *not* helpful on LeapOfthought
  - +Local Neutral term helpful on Wikidata5m
  - but adding objectives slightly lowers Update Success on Main Input

Updating Beliefs: Experiments

- Updates improve belief consistency!

Metric	Before Update	After Update
Entailment Acc	58.30 (5.7)*	75.50 (4.3)
Para. Cons (zsRE)	61.39 (1.3)	94.53 (0.6)
Para. Cons (Wiki)	24.69 (0.5)	84.56 (0.9)

Table 7: Entailment Acc and Paraphrase Consistency before and after model updates to incorrect points. *All Main Inputs in this subset are wrongly predicted as false, so the entailment does not actually hold.

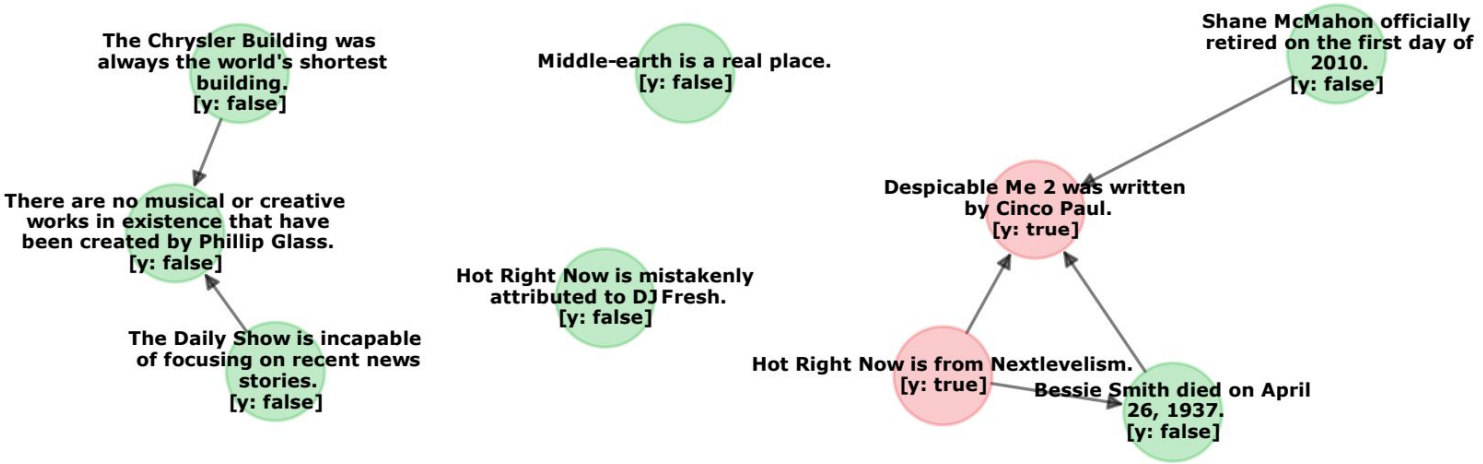
Updating Beliefs: Conclusions

- About the problem:
  - When $r_{test} = 1$, high update success but model performance typically falls on other data
  - Updates generalize across paraphrases surprisingly well
  - Sequential updating much, much harder than single-update setting
  - Retaining predictions on *Local Neutral* data harder than on *Random* data
  - Additional objective terms helpful in some settings but not always needed
  - Belief updates improve model consistency!
- About which methods are best:
  - For $r_{test} = 1$, off-the-shelf optimizers are competitive with learned optimizers
  - For $r_{test} > 1$ or seq2seq, SLAG objective greatly improves performance over KE
  - For $r_{test} > 1$ and binary tasks: off-the-shelf optimizers are best
  - For $r_{test} > 1$ and seq2seq: learned optimizers are best

Visualizing Beliefs: Approach

- Let's look at the connections between beliefs
- Beliefs are *connected* when changing one leads the other to change
  - Update belief A → *observe* a change in belief B
- Make a *belief graph*:
  - Each data point (belief) is a node
  - Edge from *u* to *v* means that changing *u* leads to change in *v*
  - We'll color nodes green when (original) model prediction is correct
- Let's look at an example for FEVER
  - Will use AdamW as update method, following experimental results
  - Update success is 100%, retain rate is 98.8%
  - Will show a non-random subgraph for illustration

Visualizing Beliefs: Experiments



Visualizing Beliefs: Experiments

- More quantitative summary:
  - # Nodes
  - % Edgeless
  - # Edges
  - # In-Edges at 95th percentile
  - # Out-Edges at 95th percentile
  - # Corrupted at 95th percentile (# predictions turned incorrect by an update)
  - Update-Transitivity (if changing A changes B and changing B changes C, does changing A change C?)

Visualizing Beliefs: Experiments

- More quantitative summary:
  - All beliefs are connected to some other belief
  - 5% of beliefs highly interconnected
  - 5% of updates highly damaging
  - Limited logical consistency

	Dataset	
Metric	FEVER	LeapOfThought
# Nodes	10,444	8,642
% Edgeless	0.0	0.0
# Edges Total	1.88m	9.71m
# In Edges (95th perc.)	1,088	5,347
# Out Edges (95th perc.)	390	3,087
# Corrupted (95th perc.)	211	2,752
% Update-Transitivity	66.64	24.38*

Visualizing Beliefs: Conclusions

- Hard to understand why individual beliefs are connected
- Some beliefs are extremely interconnected
- Models display limited logical consistency under updating
    - According to update methods we have currently

Recap

- *Detecting* beliefs (to know what they are)
  - Look for structural properties of beliefs: logical consistency and expected invariances
- *Updating* beliefs (to make them more truthful)
  - Train a learned optimizer to do the updating, encode all our goals in its objective
- *Visualizing* beliefs (to understand connections between them)
  - Look at the graph of model beliefs. Is its “worldview” reasonable?

Connections to Explainable NLP

- Normally I work on explainability
- Belief graphs could eventually be a good way to understand a model
- Ideally, we would explain model behavior *in terms of beliefs*
  - Behavior = beliefs + motivations
- Explainability is also often motivated by the promise of debugging models
- Often this looks like:
  - Detect bugs via explanations
  - Fix model by…finetuning on better data
- Sometimes explanations simplify the “finetune on better data” step
- But would be nice to have a tool that fixes bug automatically

Thank You!

Code: https://github.com/peterbhase/SLAG-Belief-Updating

Contact Info:

Peter Hase, UNC Chapel Hill
peter@cs.unc.edu
https://peterbhase.github.io

Appendix

- We want to assess the *structural properties* of model outputs
  - Are they consistent under paraphrase? **Paraphrase Consistency** (Elazar et al., 2021)
    - Fraction of all pairs of paraphrased inputs that yield the same model output
  - Are they logically consistent? **Entailment Consistency** (Talmor et al., 2020)
    - When “A is true” implies “B is true”...
    - and when models predicts A is true, how often is B predicted as True?
  - Does changing one belief correctly change other entailed beliefs?
    - When “A is true” implies “B is true”...
    - and we change model prediction on A from false to true, will model predict B is true?
  - Does changing one belief erroneously change other unrelated beliefs?
    - When “A is true” *does not* imply “B is true” or “B is false”
    - and we change model prediction on A from false to true, does model prediction for B change?

Appendix

- Hypernetwork takes model gradient as input, yields a new “gradient”
- But gradient is $d \times d$ for a square layer
- So a dense linear layer on this input would have $O(d^4)$ parameters
- Can get down to $O(2d)$ with rank-1 weight matrix and element-wise multiplication

$$
\begin{aligned}
h &= \text{LSTM}([x; \hat{y}; y^*]) \\
\{u, v, \gamma, \delta\} &= \{\text{MLP}_i(h)\}_{i=1}^4 \\
A &= \text{softmax}(u)v^T \\
B &= \text{softmax}(\gamma)\delta^T \\
\eta &= \sigma(\text{MLP}(h)) \\
\theta^* &= \theta + \eta(A \circ \nabla_\theta \mathcal{L}(x_i, y_i^*) + B)
\end{aligned}
$$

1. Embed requested update
2. Get factors for low-rank weight
3. Make low-rank weights A, B
4. Make a scaling factor
5. Apply weights element-wise to get new gradient

Appendix

Relation	% Test Data
Place of Birth	11.00
Award Received	11.00
Cause of Death	5.66
Place of Death	11.00
Place of Burial	8.33
Educated At	11.00
Child	11.00
Occupation	11.00
Spouse	11.00
Sibling	9.01

Table 11: Wikidata relations and their proportion of the test data.

Appendix

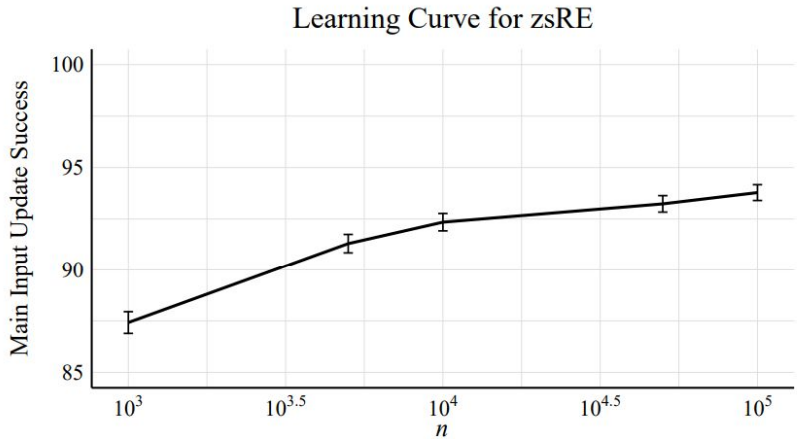


Figure 8: Main Input Update Success Rate across training set sizes, using SLAG on zsRE.

Appendix

Objective Term Ablation		Update Success Rate			Retain Predictions		Δ Acc
Dataset	Objective	Main Input	Paraphrases	Entailed Data	Local Neutral	All Data	All Data
FEVER	Main	100 (0.0)	-	-	-	98.27 (0.1)	-0.15 (0.1)
	(no KL)	100 (0.0)	-	-	-	40.42 (0.6)	-27.19 (1.2)
LeapOfThought	Main	100 (0.0)	-	76.43 (5.3)	-	96.84 (0.3)	-1.22 (0.8)
	+Ent	100 (0.0)	-	71.87 (5.3)	-	96.52 (0.3)	-0.40 (0.8)
zsRE	Main	94.46 (0.4)	94.44 (0.7)	-	-	81.96 (0.4)	-0.24 (0.1)
	+Para	93.75 (0.4)	94.41 (0.7)	-	-	75.24 (0.5)	-0.42 (0.2)
Wikidata5m	Main	88.67 (0.7)	64.12 (0.7)	-	49.78 (1.0)	71.04 (0.5)	-1.54 (0.3)
	+Para	87.46 (0.7)	81.06 (0.7)	-	47.15 (1.0)	63.02 (0.6)	-1.55 (0.3)
	+LN	87.73 (0.7)	59.75 (0.7)	-	60.49 (1.0)	72.69 (0.6)	-1.57 (0.3)
	+Para+LN	87.02 (0.7)	81.18 (0.7)	-	56.86 (1.0)	68.42 (0.6)	-1.65 (0.3)