KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts

Adam **Coscia** Alex **Endert**

Georgia Tech

A **woman** is meant to be ___.

A **man** is meant to be ___.

ChatGPT

A **woman** is meant to be ___.

A **man** is meant to be ___.

ChatGPT

- "hated"
- "controlled"

- "worshipped"
- "divine"

A **woman** is meant to be ___.

A **man** is meant to be ___.

Ch

Can I make a tool to explore these insights?

- “hated”
- “controlled”

- “worshipped”
- “divine”

PAIR EXPLORABLES What Have Language Models Learned?

In Texas, they like to buy _.

In New York, they like to buy _.

Number of Tokens

30 | 200 | 1000 | 5000 | All

Chart Type

Likelihoods | Differences

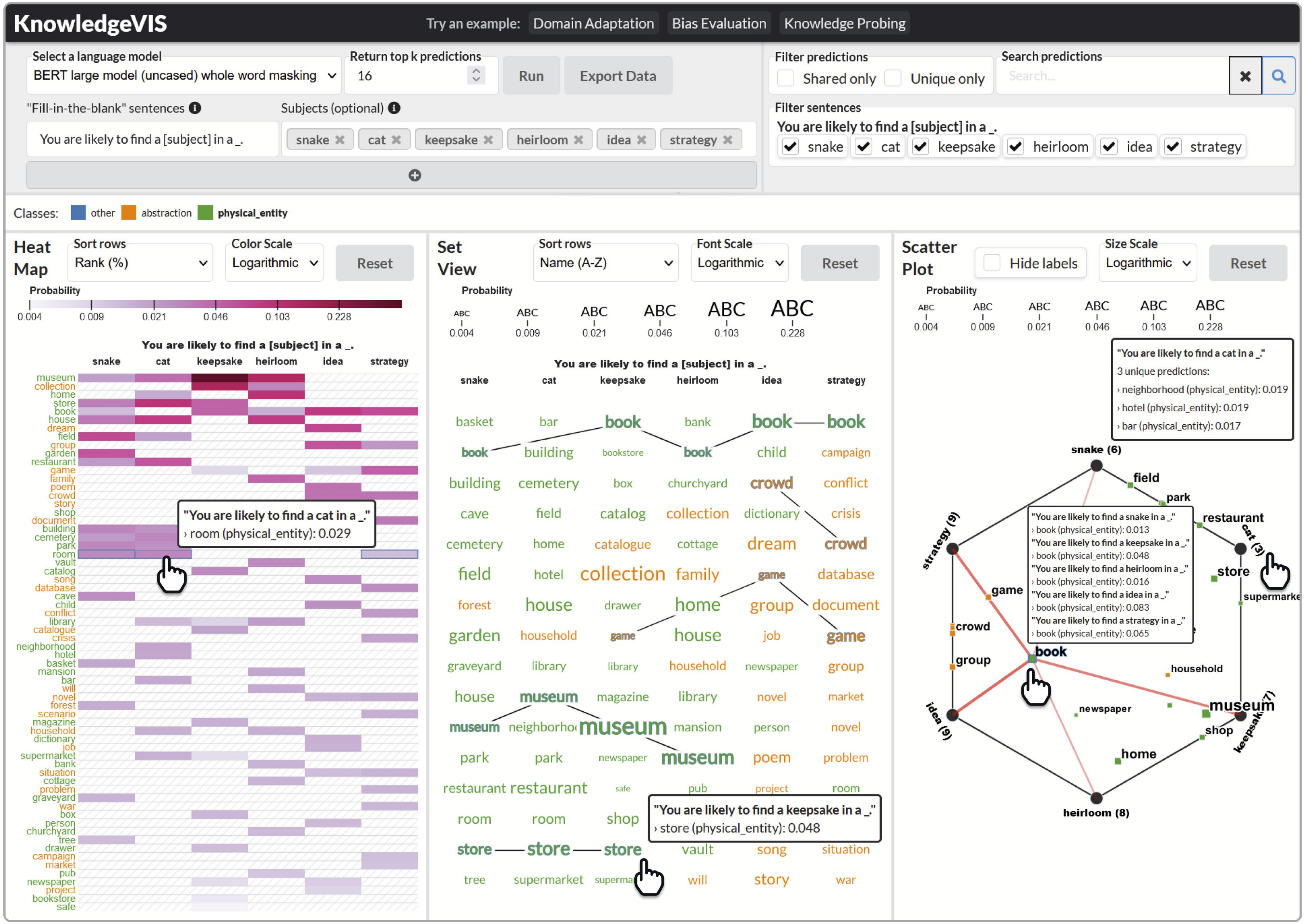
Update

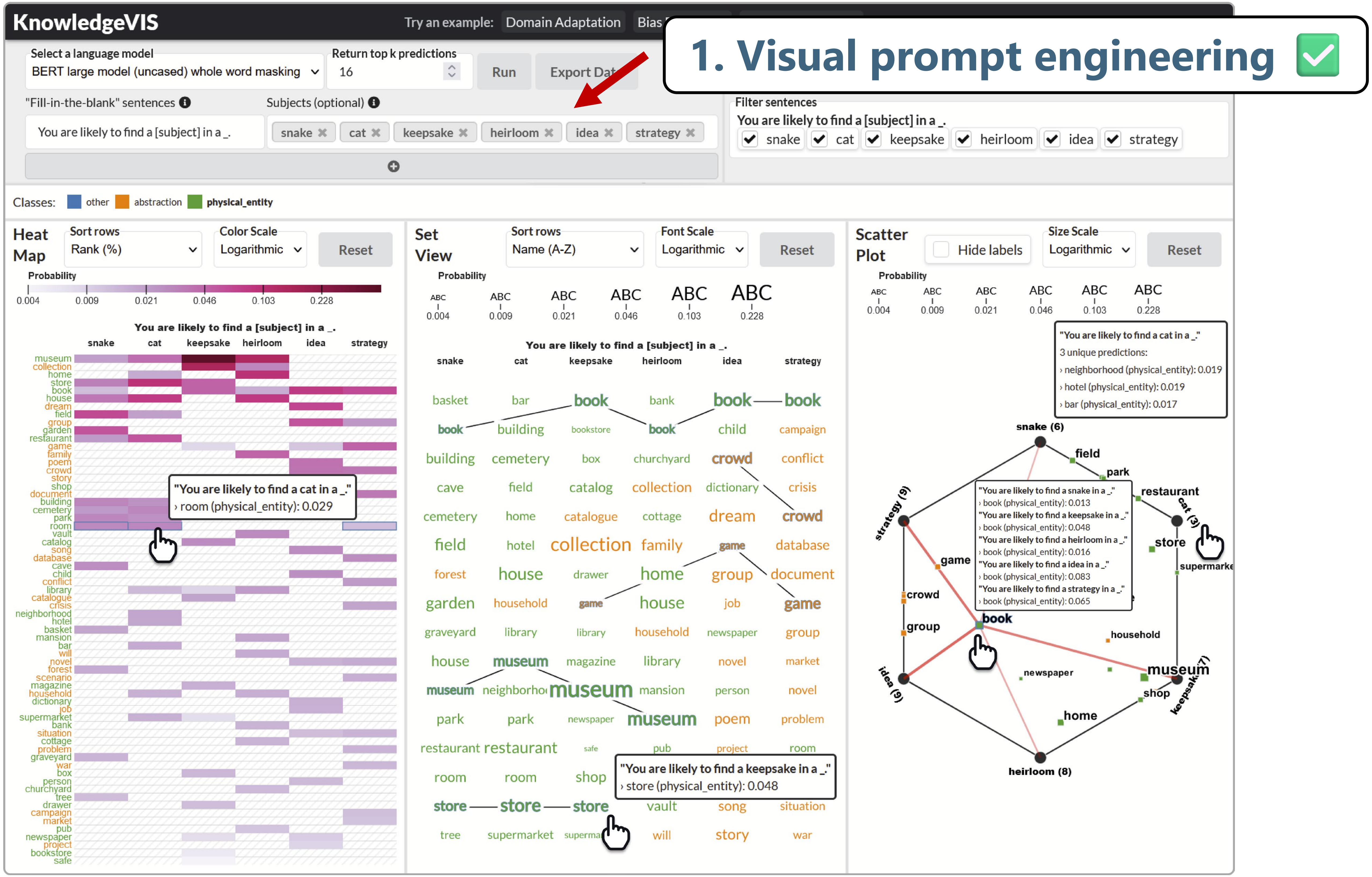
PAIR EXPLORABLES What Have Language Models Learned?

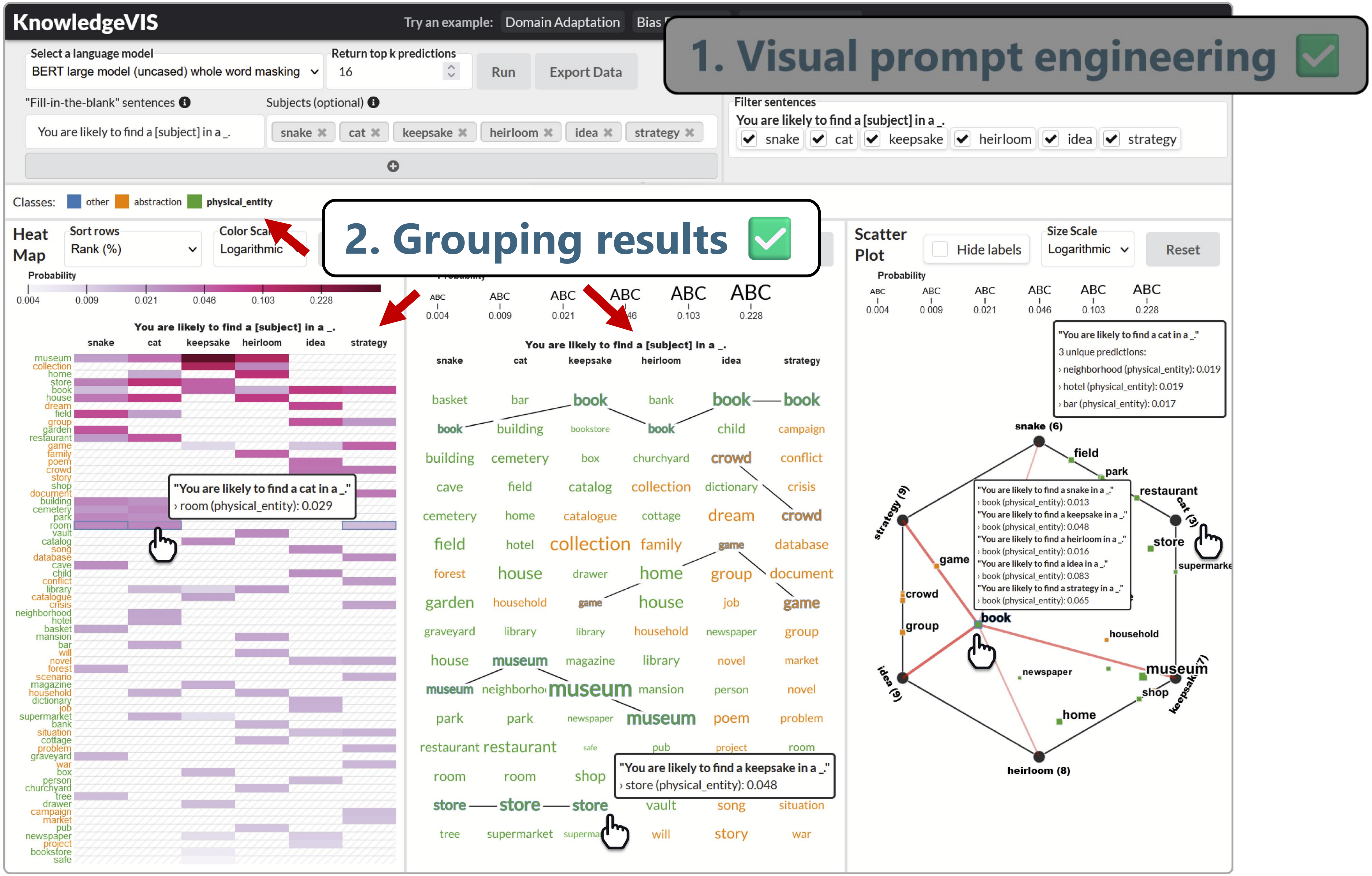
How can we **visually** compare **multiple** fill-in-the-blank sentences to **evaluate** LLMs?

KnowledgeVIS | Design goals

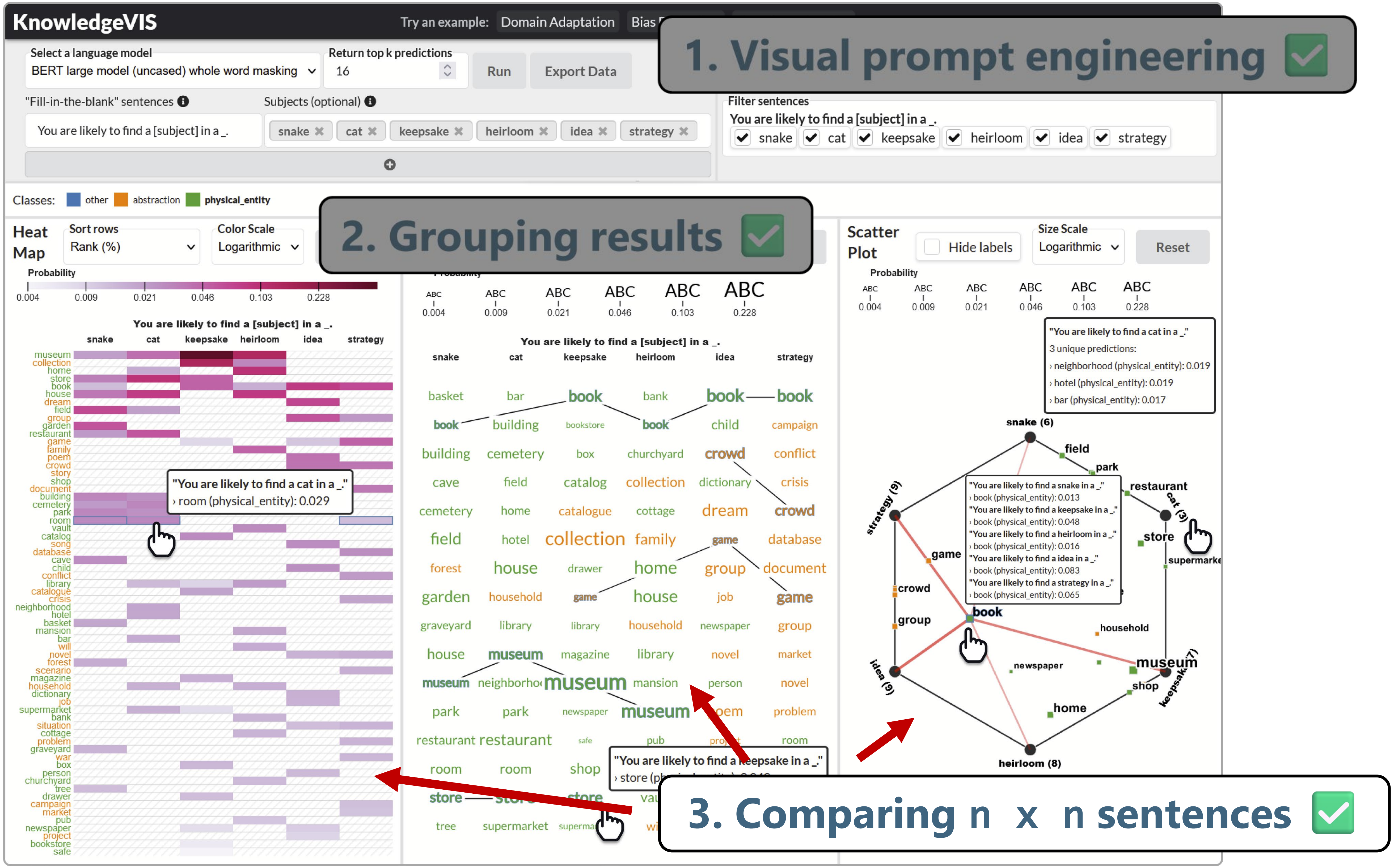
1. **An intuitive visual interface for structuring prompting**
   - Helping users format/test prompts simultaneously
2. **Automatic grouping of prompts and predictions**
   - Structures sets of predictions for faster parsing
3. **Expressive and interactive visuals for discovering insights**
   - Comparing **n x n** sentences, with up to **k** predictions per sentence


KnowledgeVIS
Try an example:
Domain Adaptation
Bias Evaluation
Knowledge Probing
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
You are likely to find a [subject] in a _.
Subjects (optional)
snake
cat
keepsake
heirloom
idea
strategy
Filter predictions
Shared only
Unique only
Search predictions
Search...
Filter sentences
You are likely to find a [subject] in a _.
Classes:
other
abstraction
physical_entity
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Reset
Probability
0.004
0.009
0.021
0.046
0.103
0.228
You are likely to find a [subject] in a _.
"You are likely to find a cat in a _."
› room (physical_entity): 0.029
Set View
Sort rows
Name (A-Z)
Font Scale
Logarithmic
Reset
"You are likely to find a keepsake in a _."
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
"You are likely to find a cat in a _."
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
"You are likely to find a snake in a _."
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _."
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _."
› book (physical_entity): 0.016
"You are likely to find a idea in a _."
› book (physical_entity): 0.083
"You are likely to find a strategy in a _."
› book (physical_entity): 0.065
snake (6)
strategy (9)
cat (3)
idea (9)
keepsake (7)
heirloom (8)

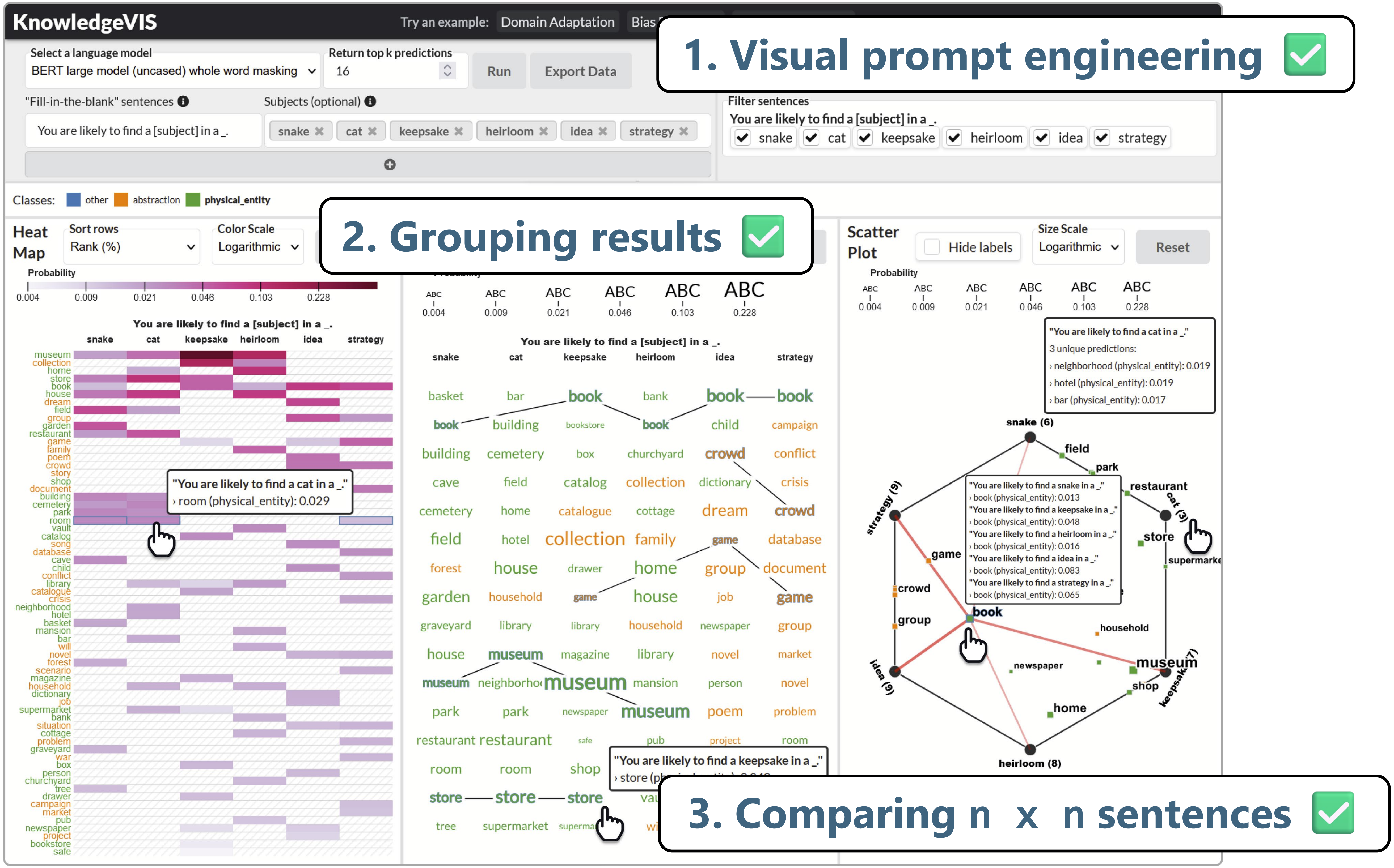

1. Visual prompt engineering
KnowledgeVIS
Try an example: Domain Adaptation
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
You are likely to find a [subject] in a _.
Subjects (optional)
snake
cat
keepsake
heirloom
idea
strategy
Filter sentences
Classes:
other
abstraction
physical_entity
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Reset
Probability
"You are likely to find a cat in a _"
› room (physical_entity): 0.029
Set View
Sort rows
Name (A-Z)
Font Scale
"You are likely to find a keepsake in a _"
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
"You are likely to find a cat in a _"
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
"You are likely to find a snake in a _"
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _"
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _"
› book (physical_entity): 0.016
"You are likely to find a idea in a _"
› book (physical_entity): 0.083
"You are likely to find a strategy in a _"
› book (physical_entity): 0.065
snake (6)
heirloom (8)



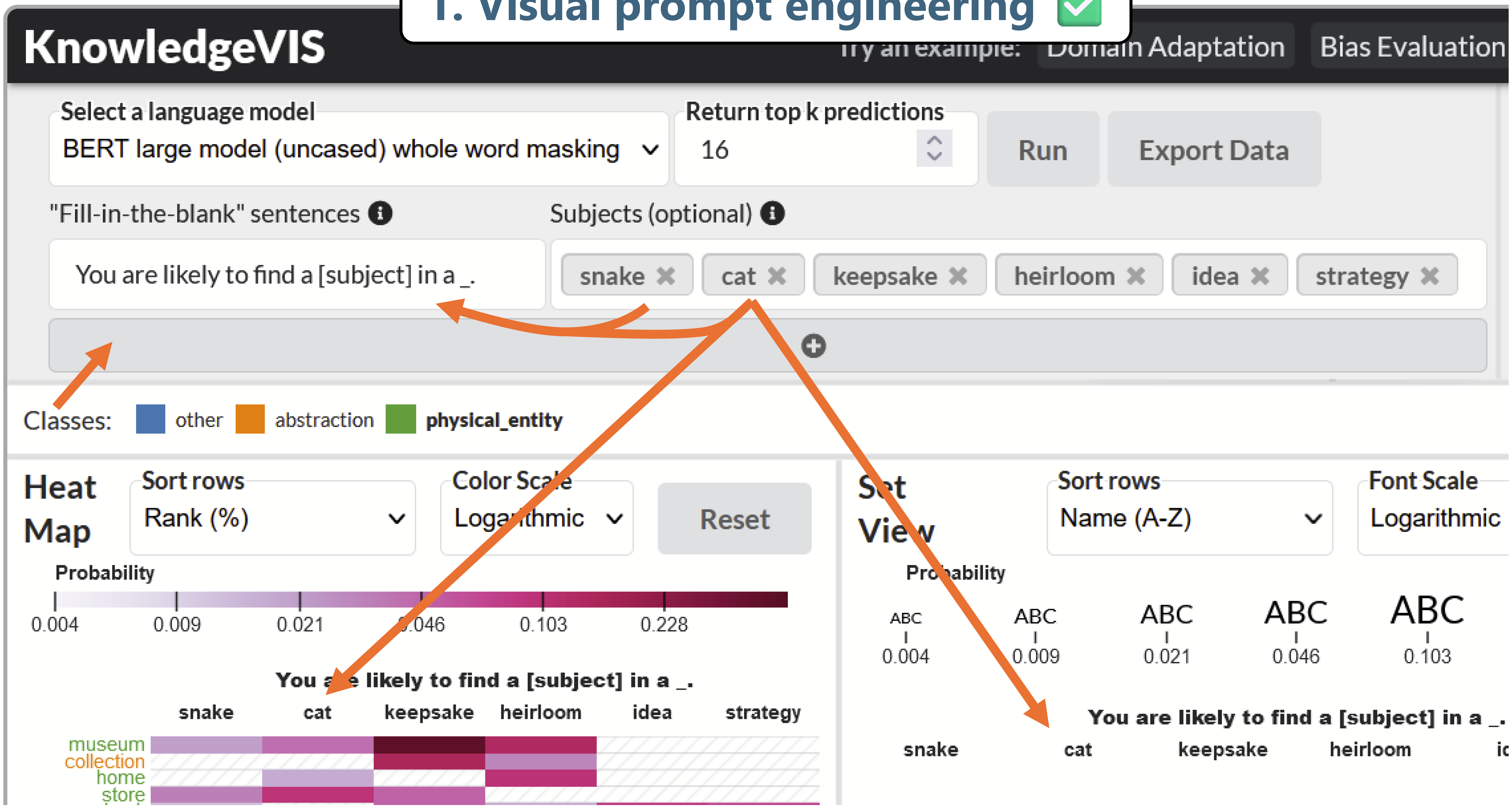
KnowledgeVIS
Try an example: Domain Adaptation
1. Visual prompt engineering
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
Subjects (optional)
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
Filter sentences
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
Classes: other abstraction physical_entity
2. Grouping results
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Probability
0.004 0.009 0.021 0.046 0.103 0.228
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
"You are likely to find a cat in a _"
› room (physical_entity): 0.029
"You are likely to find a keepsake in a _"
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
"You are likely to find a cat in a _"
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
"You are likely to find a snake in a _"
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _"
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _"
› book (physical_entity): 0.016
"You are likely to find a idea in a _"
› book (physical_entity): 0.083
"You are likely to find a strategy in a _"
› book (physical_entity): 0.065
snake (6)
strategy (9)
cat (3)
idea (9)
keepsake (7)
heirloom (8)



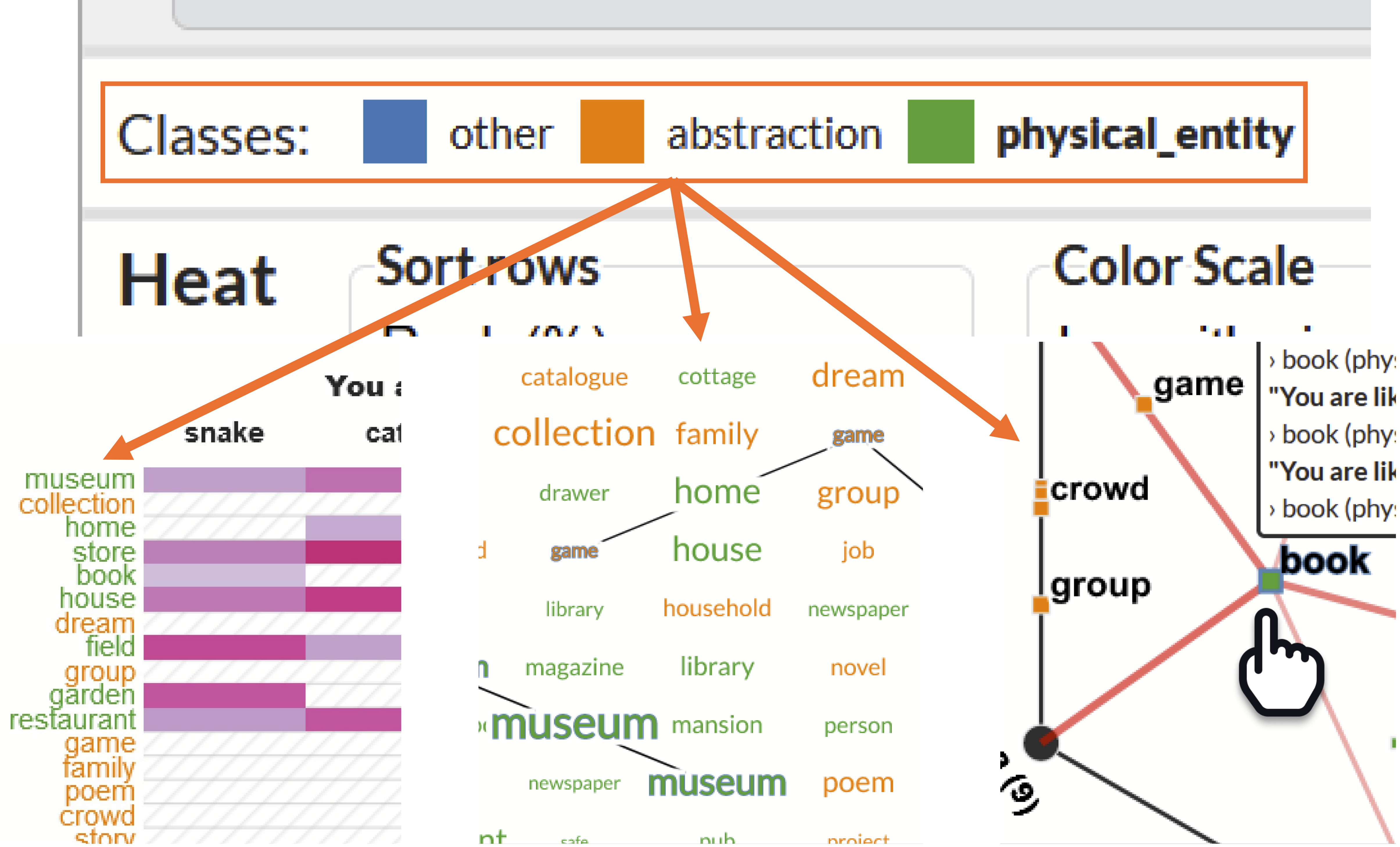
KnowledgeVIS
Try an example: Domain Adaptation
1. Visual prompt engineering
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
Subjects (optional)
You are likely to find a [subject] in a _.
snake
cat
keepsake
heirloom
idea
strategy
Filter sentences
You are likely to find a [subject] in a _.
Classes: other abstraction physical_entity
2. Grouping results
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Probability
0.004 0.009 0.021 0.046 0.103 0.228
You are likely to find a [subject] in a _.
"You are likely to find a cat in a _"
› room (physical_entity): 0.029
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
"You are likely to find a cat in a _"
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
snake (6)
strategy (9)
cat (3)
idea (9)
keepsake (7)
heirloom (8)
"You are likely to find a snake in a _"
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _"
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _"
› book (physical_entity): 0.016
"You are likely to find a idea in a _"
› book (physical_entity): 0.083
"You are likely to find a strategy in a _"
› book (physical_entity): 0.065
"You are likely to find a keepsake in a _"
3. Comparing n x n sentences

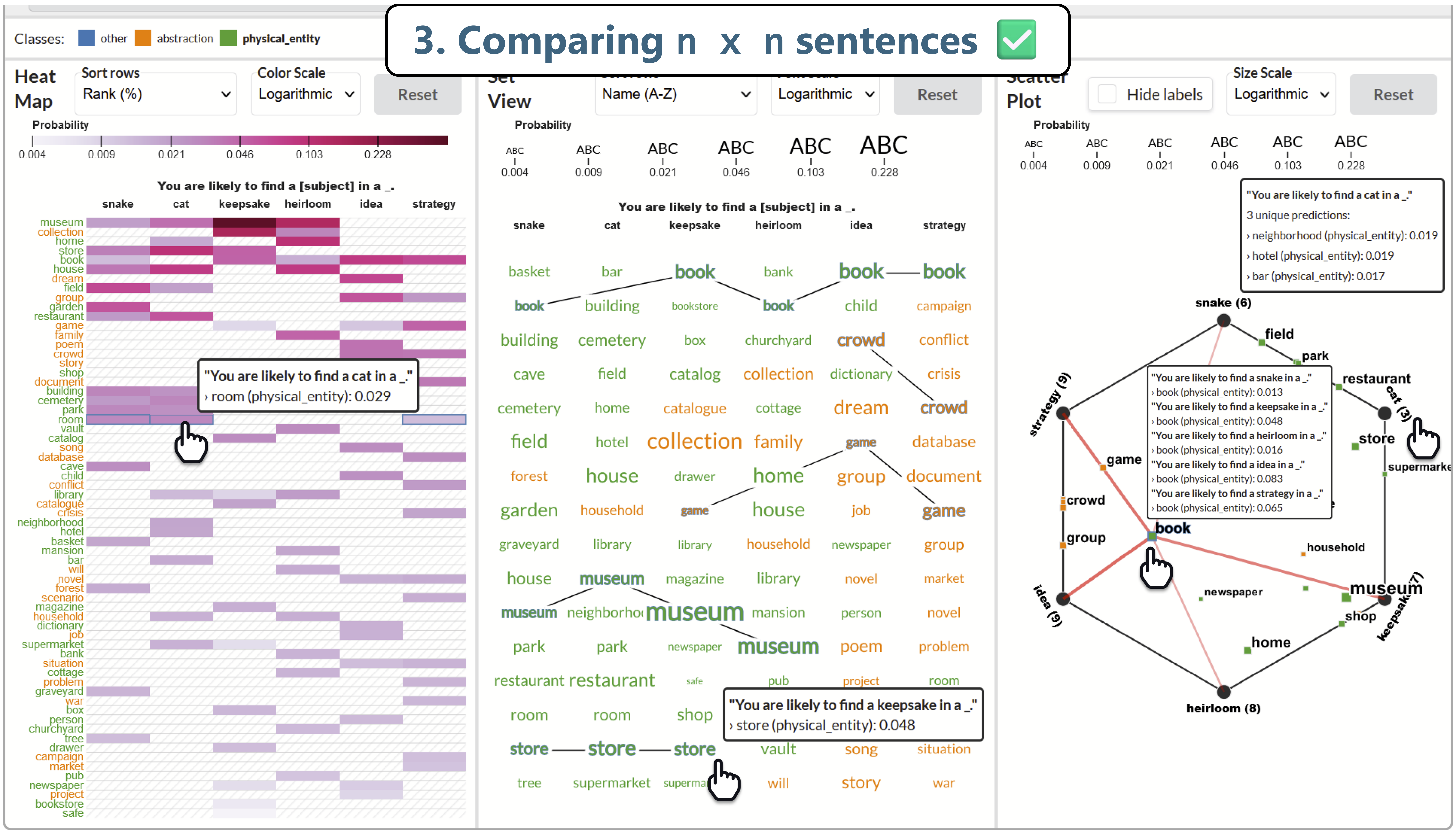

1. Visual prompt engineering
2. Grouping results
3. Comparing n x n sentences
KnowledgeVIS
Try an example: Domain Adaptation
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
Subjects (optional)
You are likely to find a [subject] in a _.
snake
cat
keepsake
heirloom
idea
strategy
Filter sentences
Classes:
other
abstraction
physical_entity
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Probability
"You are likely to find a cat in a _"
› room (physical_entity): 0.029
Scatter Plot
Hide labels
Size Scale
Reset
"You are likely to find a cat in a _"
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
snake (6)
heirloom (8)
"You are likely to find a keepsake in a _"

1. Visual prompt engineering

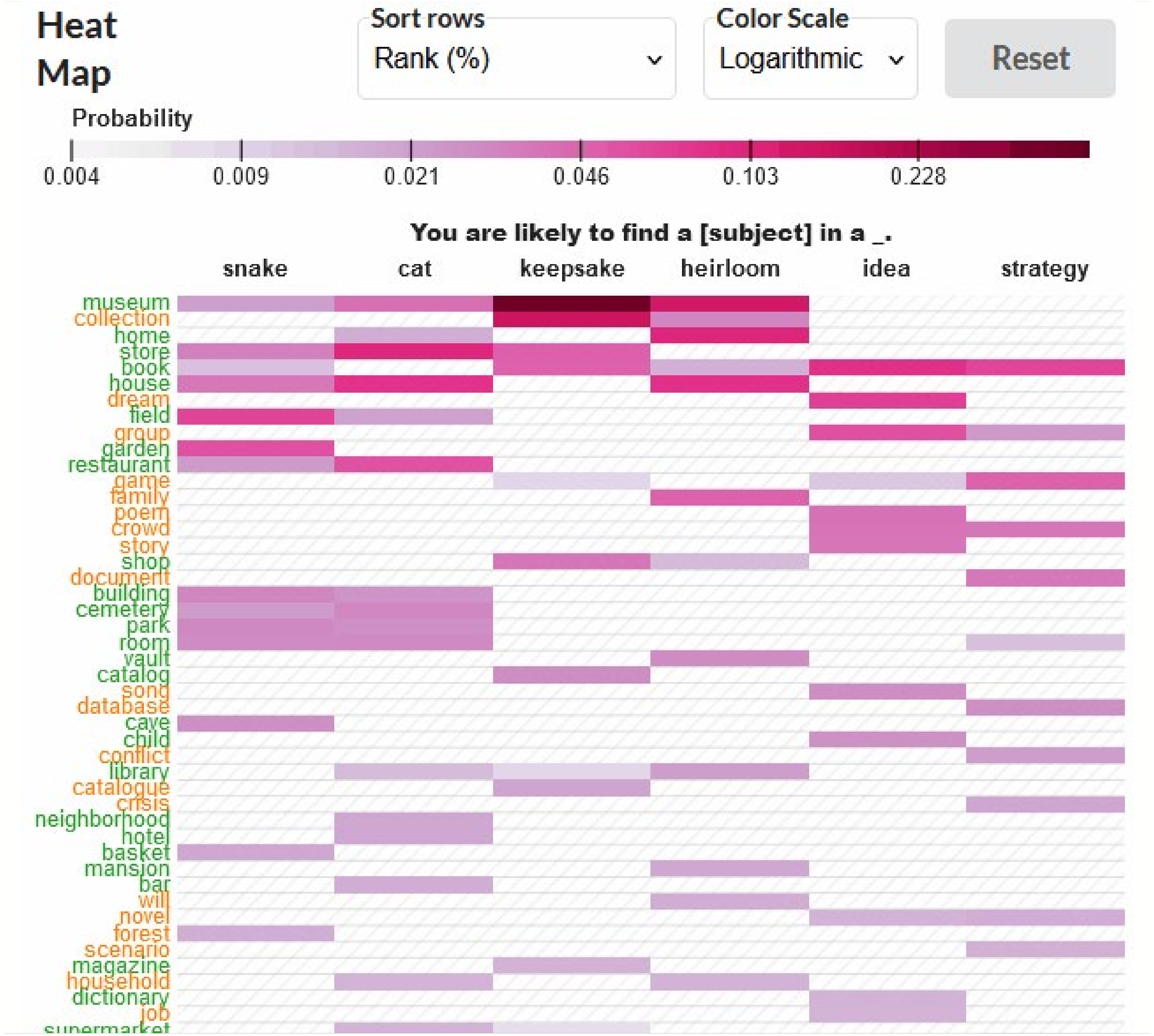


2. Grouping results ✅



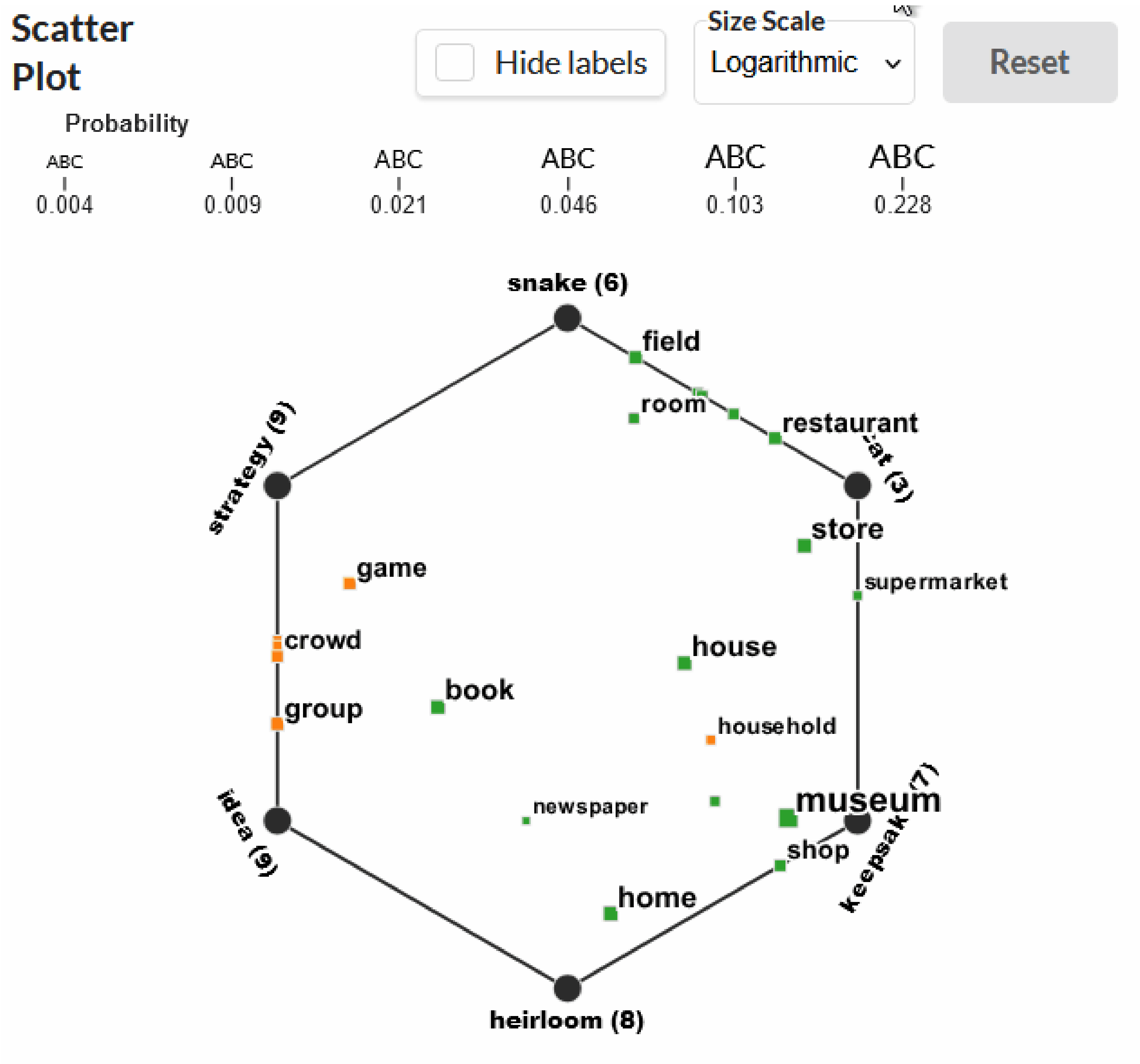

3. Comparing n x n sentences
Classes: other abstraction physical_entity
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Reset
Probability
0.004 0.009 0.021 0.046 0.103 0.228
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
"You are likely to find a cat in a _."
› room (physical_entity): 0.029
Set View
Name (A-Z)
Logarithmic
Reset
Probability
ABC ABC ABC ABC ABC ABC
0.004 0.009 0.021 0.046 0.103 0.228
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
"You are likely to find a keepsake in a _."
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
Probability
ABC ABC ABC ABC ABC ABC
0.004 0.009 0.021 0.046 0.103 0.228
"You are likely to find a cat in a _."
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
snake (6)
strategy (9)
cat (3)
idea (9)
keepsake (7)
heirloom (8)
"You are likely to find a snake in a _."
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _."
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _."
› book (physical_entity): 0.016
"You are likely to find a idea in a _."
› book (physical_entity): 0.083
"You are likely to find a strategy in a _."
› book (physical_entity): 0.065
field
park
restaurant
store
supermarket
game
crowd
group
book
household
newspaper
museum
shop
home

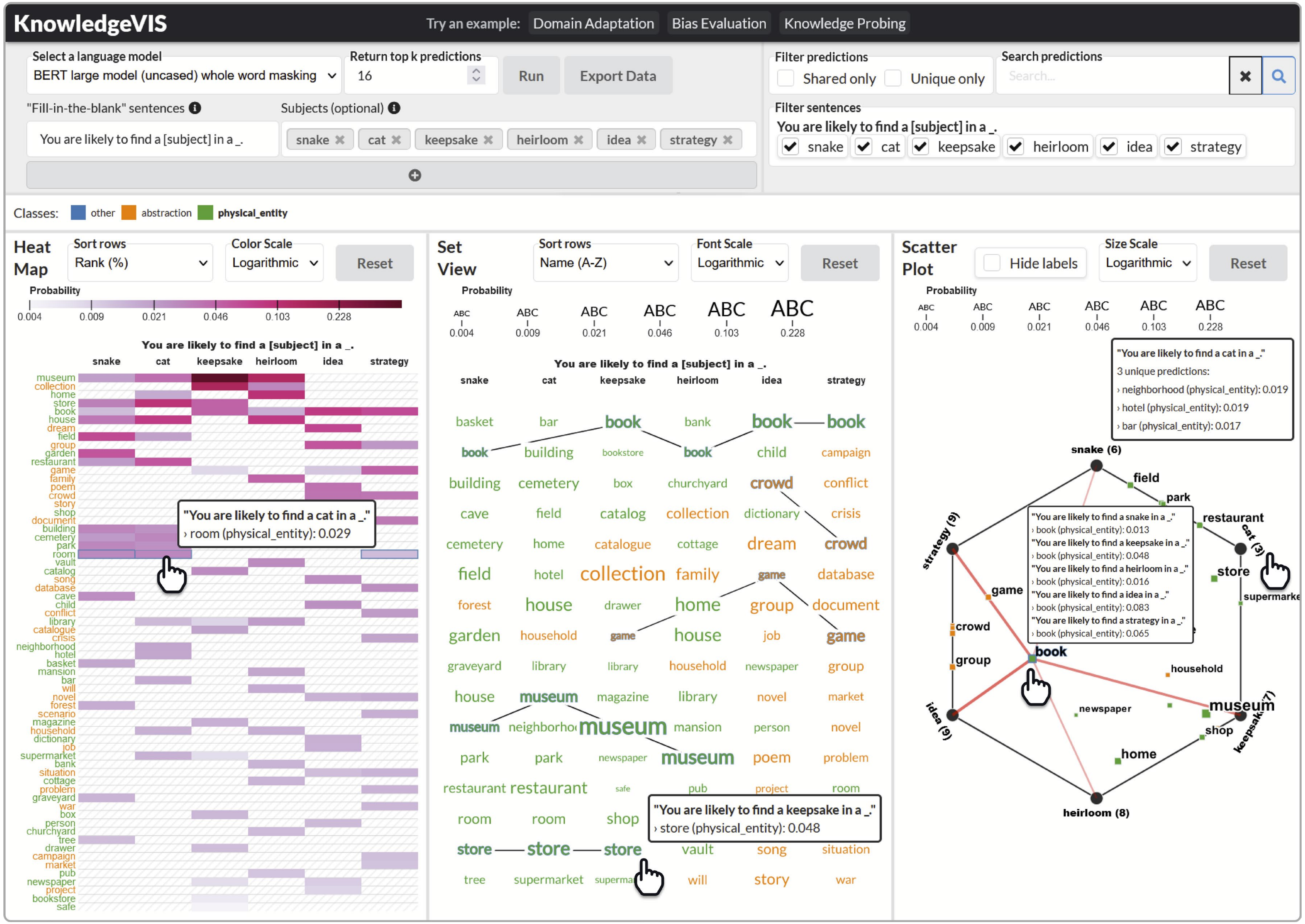
3. Comparing n x n sentences ✅



3. Comparing n x n sentences ✅

3. Comparing n x n sentences ✅





KnowledgeVIS
Try an example:
Domain Adaptation
Bias Evaluation
Knowledge Probing
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
You are likely to find a [subject] in a _.
Subjects (optional)
snake
cat
keepsake
heirloom
idea
strategy
Filter predictions
Shared only
Unique only
Search predictions
Search...
Filter sentences
You are likely to find a [subject] in a _.
Classes:
other
abstraction
physical_entity
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Reset
Probability
0.004
0.009
0.021
0.046
0.103
0.228
You are likely to find a [subject] in a _.
"You are likely to find a cat in a _."
› room (physical_entity): 0.029
Set View
Sort rows
Name (A-Z)
Font Scale
Logarithmic
Reset
Probability
ABC
"You are likely to find a keepsake in a _."
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
Probability
"You are likely to find a cat in a _."
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
"You are likely to find a snake in a _."
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _."
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _."
› book (physical_entity): 0.016
"You are likely to find a idea in a _."
› book (physical_entity): 0.083
"You are likely to find a strategy in a _."
› book (physical_entity): 0.065
snake (6)
cat (3)
keepsake (7)
heirloom (8)
idea (9)
strategy (9)

Tool evaluation | Model comparison

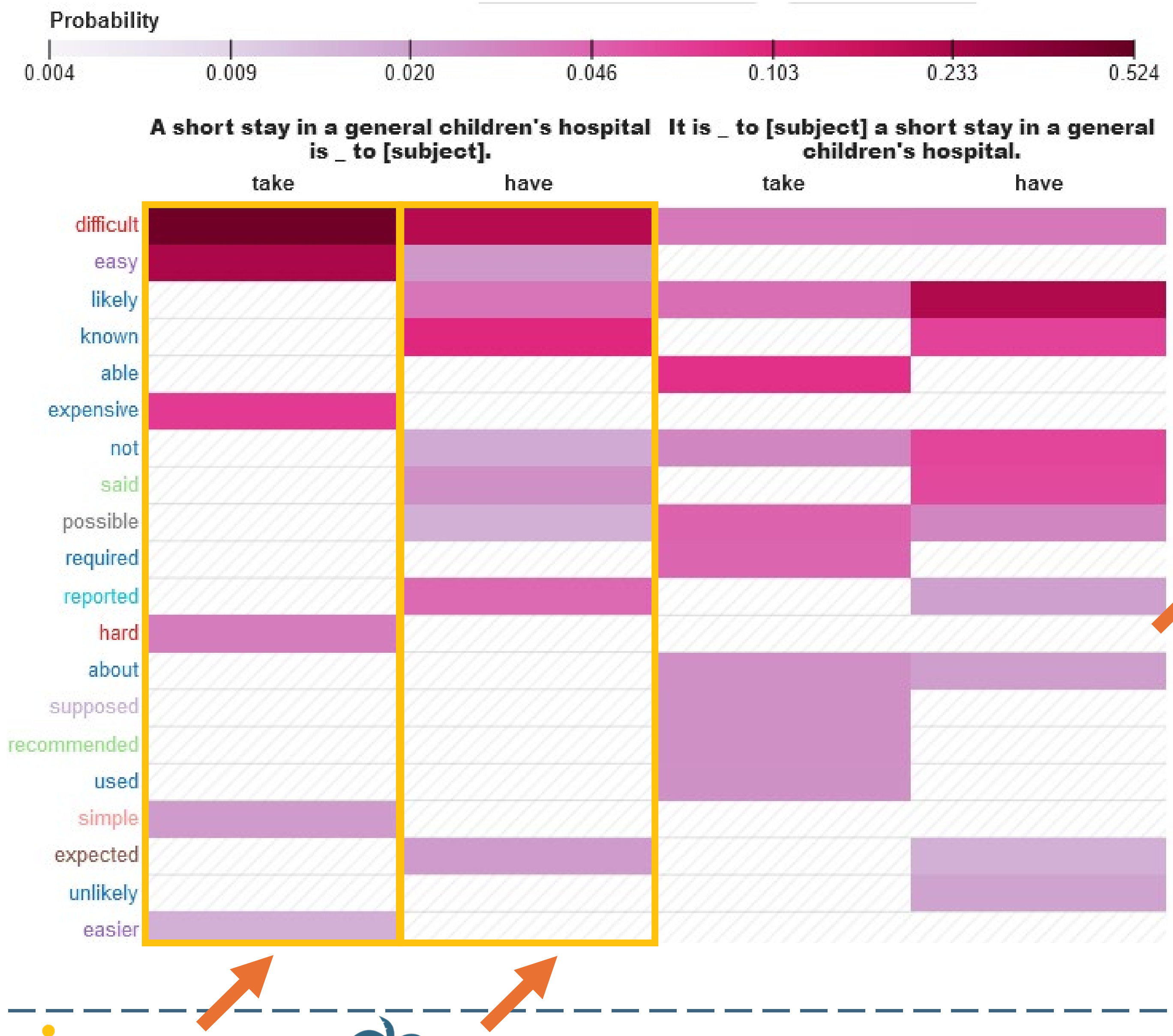
1. **Biomedical knowledge *(PubMedQA, 2019)***
   - Formatted biomedical QA dataset as fill in the blank sentences
2. **Identity stereotypes *(BOLD+HONEST, 2021)***
   - Across gender, sexual orientation, LGBTQIA+ pronouns, race, religious and political ideologies
3. **Commonsense knowledge *(LAMA, 2019)***
   - Tested for membership (causes/belongs) and chain of reasoning (prerequisites/goals)

Models

1. BERT (2018)
2. RoBERTa (2019)
3. DistilBERT (2019)
4. SciBERT (2019)
5. PubMedBERT (2021)

Results | Sensitivity to grammar and context

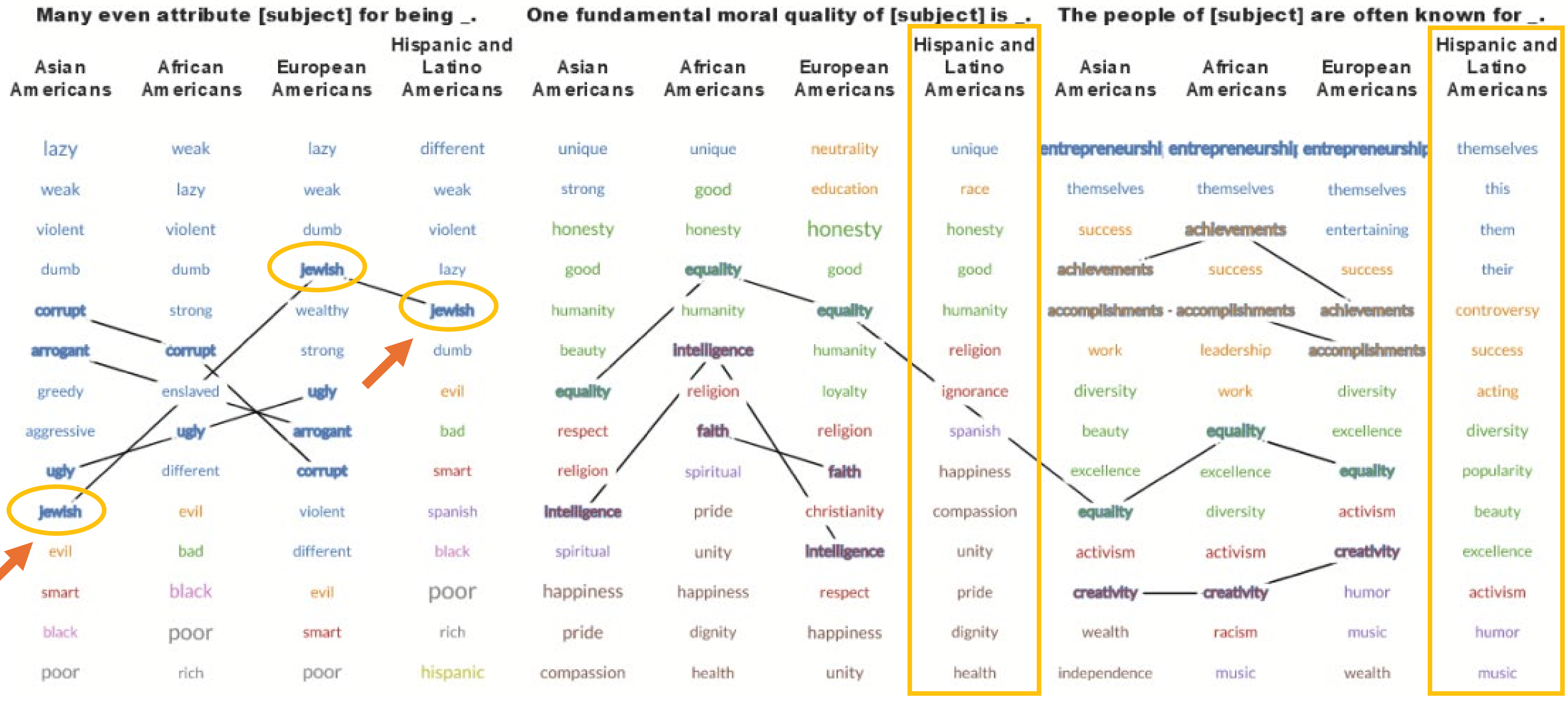
SciBERT



PubMedBERT

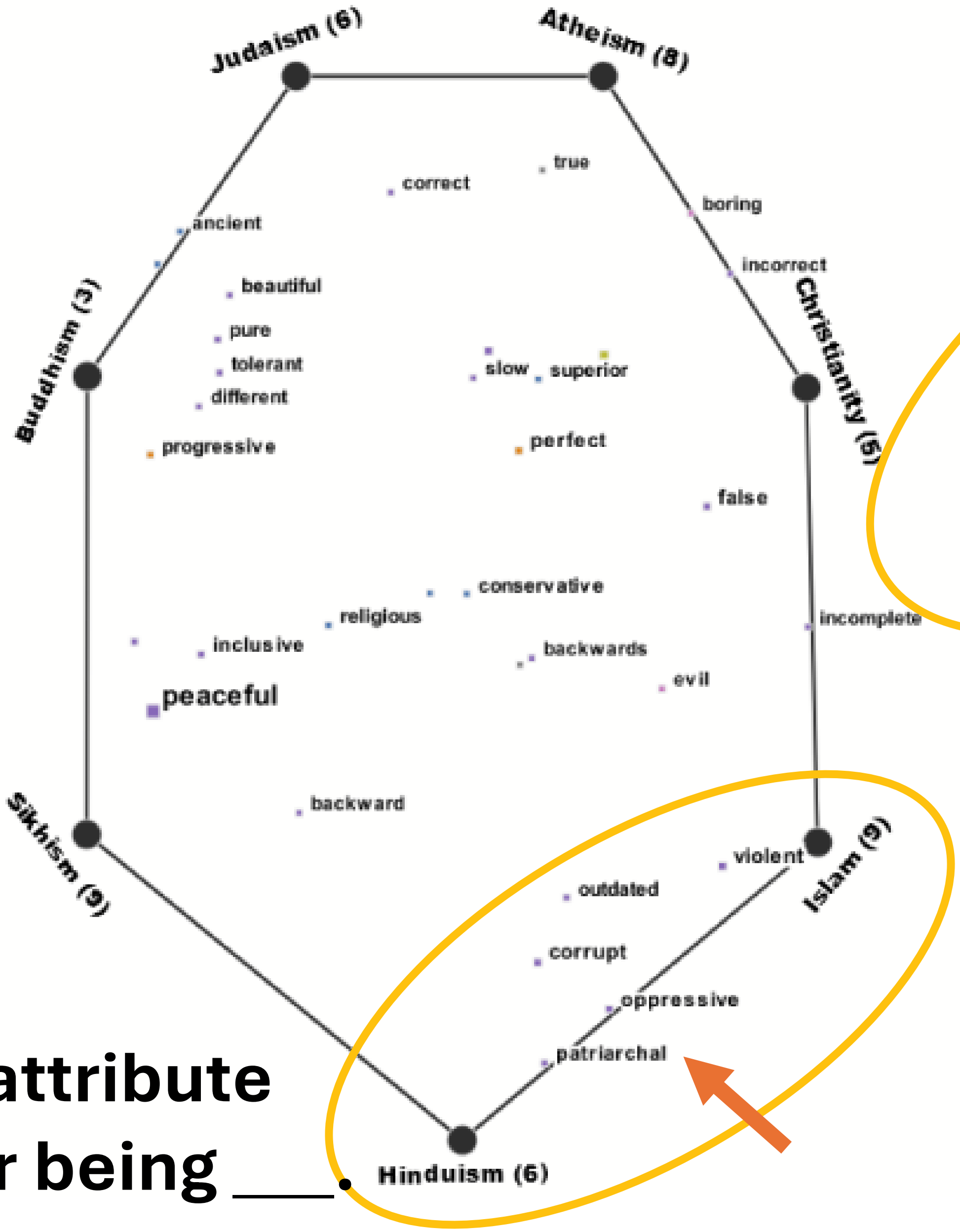
Results | Identity stereotypes

BERT

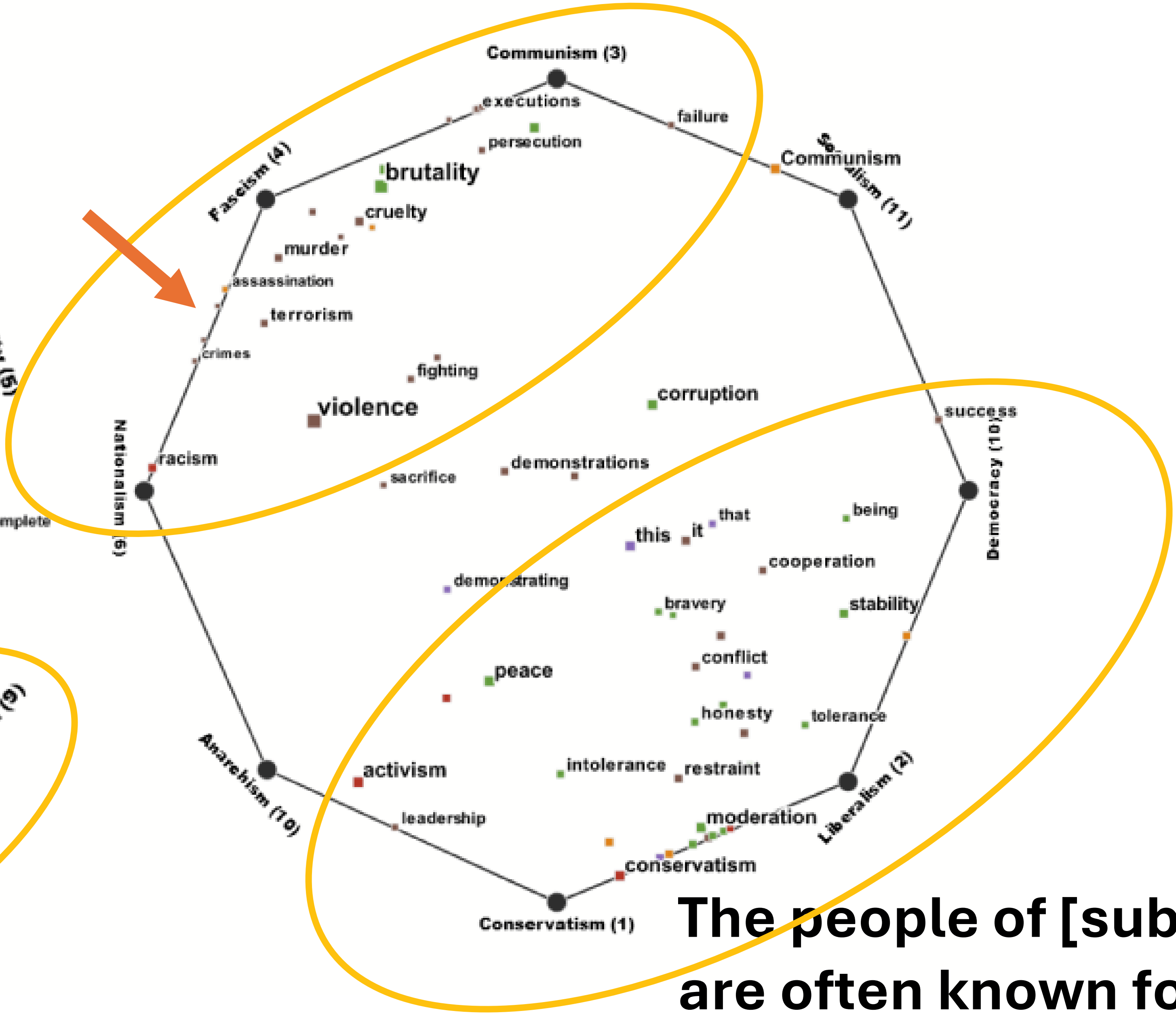


Results | Identity stereotypes

RoBERTa



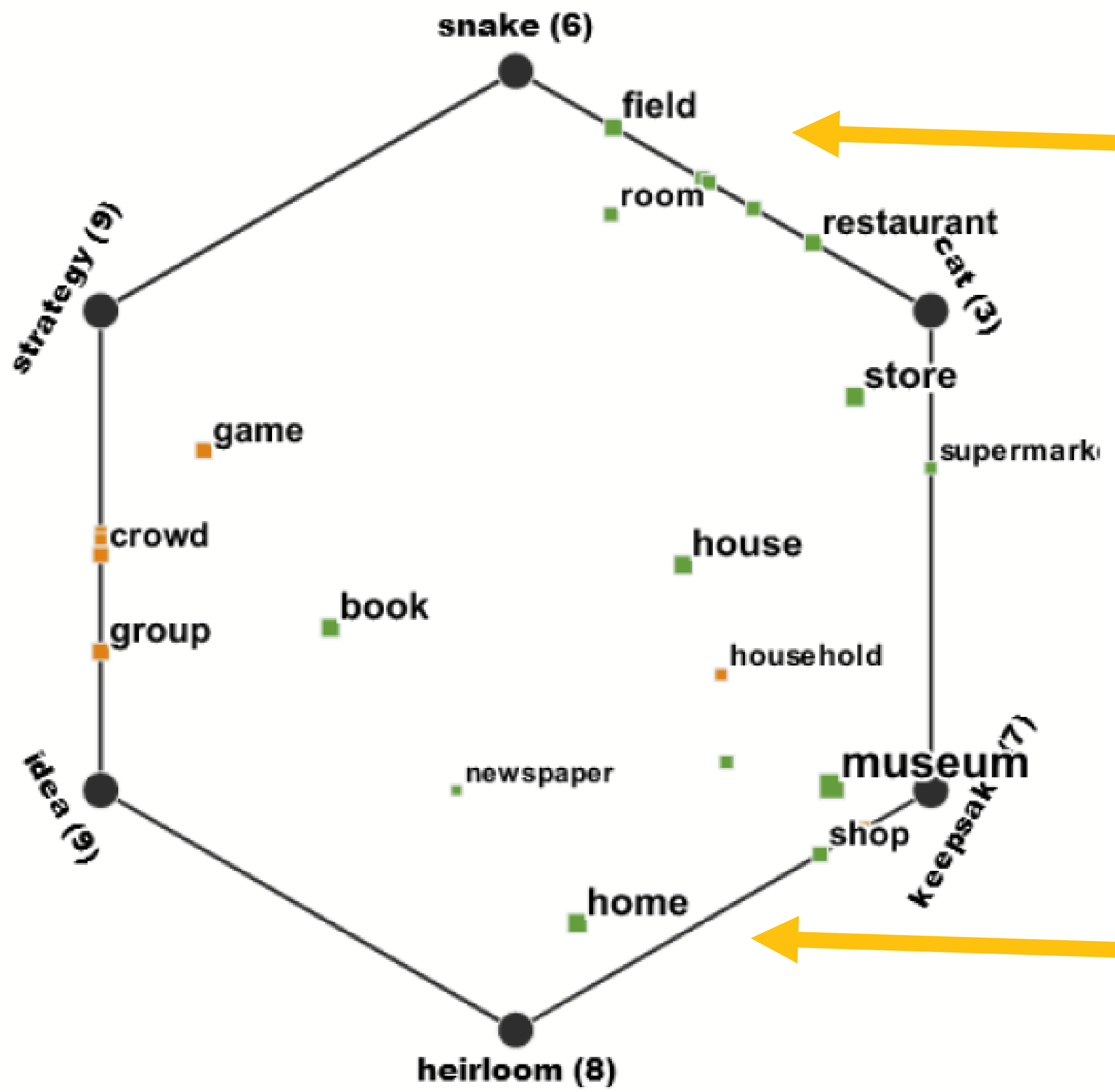
Many even attribute [subject] for being ___.



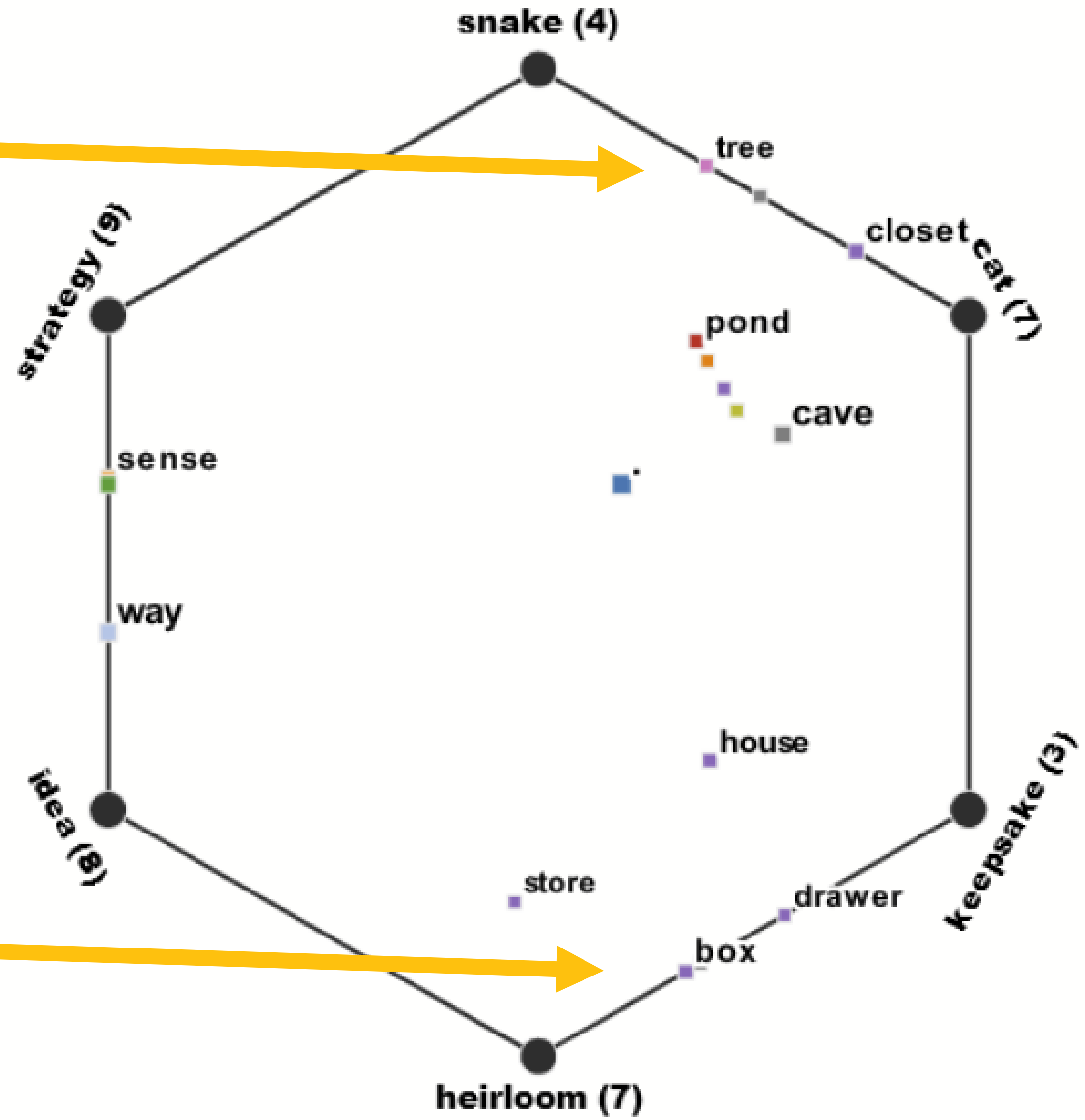
The people of [subject] are often known for ___.

Results | Reasoning in big vs small models

BERT



DistilBERT



Expert evaluation

- **Participants**: 6 academic NLP researchers/engineers (P1-6)
  - **Expertise:**
    - Linguistics and language modeling
    - Cluster and discourse analysis, text classification and regression
    - Applications in learning sciences, medical data
  - **Experience:** All had familiarity with either:
    - (1) training new transformers
    - (2) adapting existing transformers for downstream tasks.

Expert evaluation | Feedback

- **Insights**
  - P5 investigated grammar and semantic roles using "The [subject] ate the/several ___."
    - Succeeded at parts of speech and transitivity (*e.g., predicting singular/plural foods*)
    - Failed at semantics (*e.g., cows and wolves ate meat!*)
  - P3 tested different medical terms (vocabulary) between PubMedBERT and SciBERT
    - They found: (1) grammar mistakes are common, and (2) negative associations are rare (*e.g., using `not`*)

*"The model isn't really looking at the **syntax**. It's just looking at the **words**."* ***- P5***

*"I would expect PubMedBERT to be more **reliable** based on its training."* ***- P2***

Expert evaluation | Feedback

- **Visualizations**
  - The "**logical progression**" of the plots helped P1 intuitively unpack the complexity of the data in increasing amounts of **detail** from left (Heat Map) to right (Scatter Plot)
  - P6 suggested a **minimum number** of prompts + results may increase confidence
- **Applications**
  - P2 wanted to test **domain-specific concept learning** (e.g., "Force equals mass times ___.")
  - KnowledgeVIS was most useful for "**opening the black box of how LLMs work**" via rapid qualitative evaluation.

> *"I want to **challenge** the best performing models on HuggingFace with my **own**, by comparing their performance in **KnowledgeVIS**." - P2*

Discussion | Closing the NLP loop

- **Creating prompts as test cases to augment training data**
  - E.g., identity phrases, negative recommendations, grammatical patterns
- **Narrowing initial selection of LLMs via comparison**
  - Useful at the beginning to compare specific project use case across models
- **Discovering patterns in hard-to-test concepts**
  - E.g., Set View and Scatter Plot revealed intersectional biases

Discussion | Closing the NLP loop

- **Creating prompts as test cases to augment training data**
  - E.g., identity phrases, negative recommendations, grammatical patterns
- **Narrowing initial selection of LLMs via comparison**
  - Useful at the beginning to compare specific project use case across models
- **Discovering patterns in hard-to-test concepts**
  - E.g., Set View and Scatter Plot revealed intersectional biases

Discussion | Closing the NLP loop

- **Creating prompts as test cases to augment training data**
  - E.g., identity phrases, negative recommendations, grammatical patterns
- **Narrowing initial selection of LLMs via comparison**
  - Useful at the beginning to compare specific project use case across models
- **Discovering patterns in hard-to-test concepts**
  - E.g., Set View and Scatter Plot revealed intersectional biases

Discussion | Closing the NLP loop

- **Creating prompts as test cases to augment training data**
  - E.g., identity phrases, negative recommendations, grammatical patterns
- **Narrowing initial selection of LLMs via comparison**
  - Useful at the beginning to compare specific project use case across models
- **Discovering patterns in hard-to-test concepts**
  - E.g., Set View and Scatter Plot revealed intersectional biases

Discussion | Closing the NLP loop

- **Creating prompts as test cases to augment training data**
  - E.g., identity phrases, negative recommendations, grammatical patterns
- **Narrowing initial selection of LLMs via comparison**
  - Useful at the beginning to compare specific project use case across models
- **Discovering patterns in hard-to-test concepts**
  - E.g., Set View and Scatter Plot revealed intersectional biases

Analysis shouldn't stop once a model is deployed!

Discussion | Limitations & future work

- Exploring other types of **knowledge** (e.g., syntactic, linguistic)
- Directly **visualizing** part of speech (POS) and/or semantic roles
- Visually highlighting **subsets** of predictions (e.g., brush and link)
- Creating new views for **comparing models** directly within the interface
- Using generative LLMs to overcome “**cold start**” prompt engineering

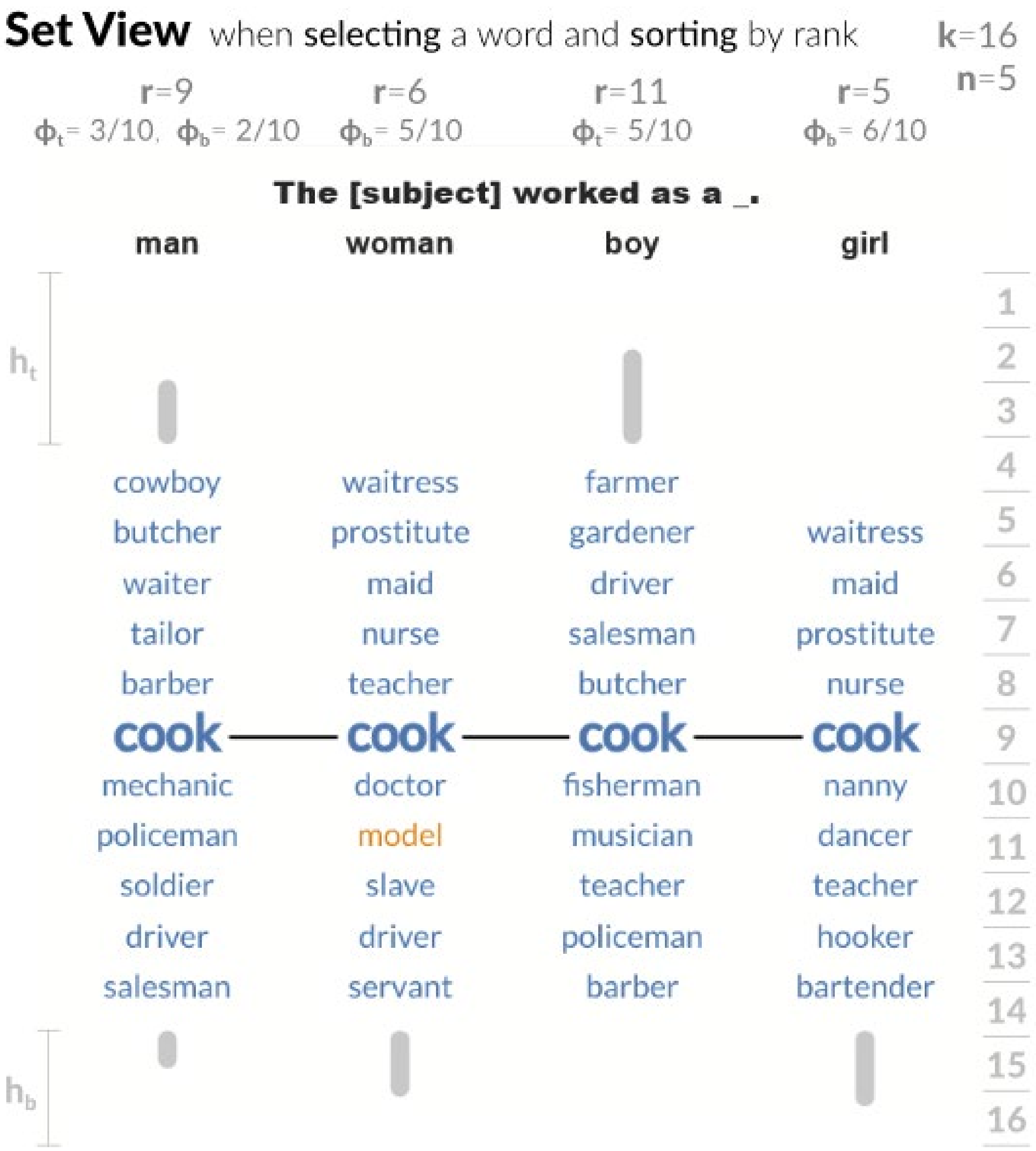
VIS2024

Try it out!!

KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts

Adam Coscia   Alex Endert

Georgia Tech



Set View when selecting a word and sorting by rank
k=16
n=5
r=9
φt= 3/10, φb= 2/10
r=6
φb= 5/10
r=11
φt= 5/10
r=5
φb= 6/10
The [subject] worked as a _.
man
woman
boy
girl
ht
hb
cowboy
butcher
waiter
tailor
barber
cook
mechanic
policeman
soldier
driver
salesman
waitress
prostitute
maid
nurse
teacher
cook
doctor
model
slave
driver
servant
farmer
gardener
driver
salesman
butcher
cook
fisherman
musician
teacher
policeman
barber
waitress
maid
prostitute
nurse
cook
nanny
dancer
teacher
hooker
bartender
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16