ALOHA: Artificial Learning of Human Attributes for Dialogue Agents

Aaron W. Li, Veronica Jiang*, **Steven Y. Feng***,
Julia Sprague, Wei Zhou, Jesse Hoey

UNIVERSITY OF WATERLOO

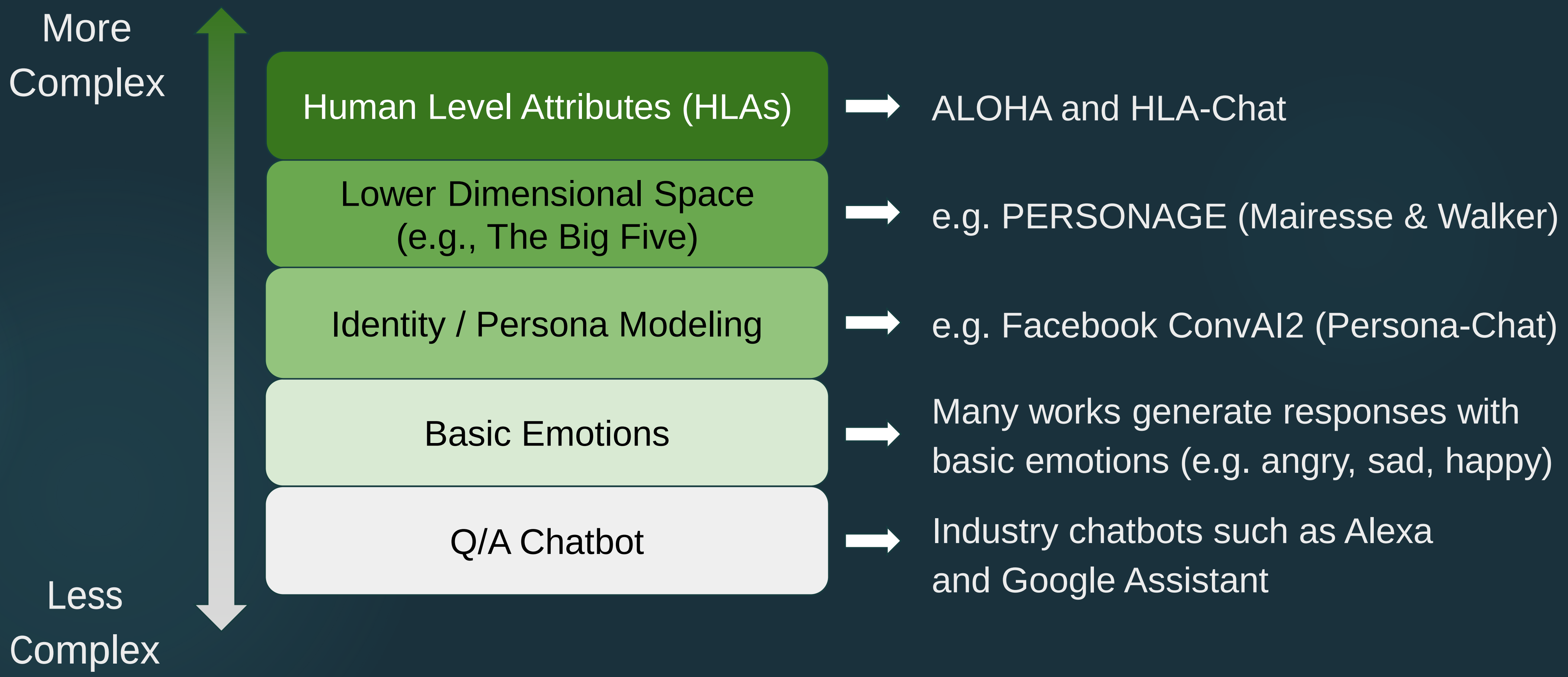
HUAWEI

Overall Goal

- Giving chatbots and virtual assistants the ability to imitate and express human emotions/personality

- How?
  - **Human-Level Attributes (HLA)** - Based on tropes: aspects of fictional characters representative of their identity
  - **HLA-Chat**: Dataset of characters with their HLAs + dialogue
  - **Artificial Learning of Human Attributes (ALOHA)**: System to retrieve character specific language models

Related Work



More Complex

Human Level Attributes (HLAs) → ALOHA and HLA-Chat

Lower Dimensional Space (e.g., The Big Five) → e.g. PERSONAGE (Mairesse & Walker)

Identity / Persona Modeling → e.g. Facebook ConvAI2 (Persona-Chat)

Basic Emotions → Many works generate responses with basic emotions (e.g. angry, sad, happy)

Q/A Chatbot → Industry chatbots such as Alexa and Google Assistant

Less Complex

Human Level Attributes (HLAs)

Broad attribute: Friendly

[HLA: Childhood Friends]

...

[HLA: Vitriolic Best Buds]

Broad attribute: Trustworthy

[HLA: The Reliable One]

...

[HLA: Only Friend]

Broad attribute: Helpful

[HLA: The Caretaker]

...

[HLA: We Help the Helpless]

Broad attribute: Curious

[HLA: The Watson]

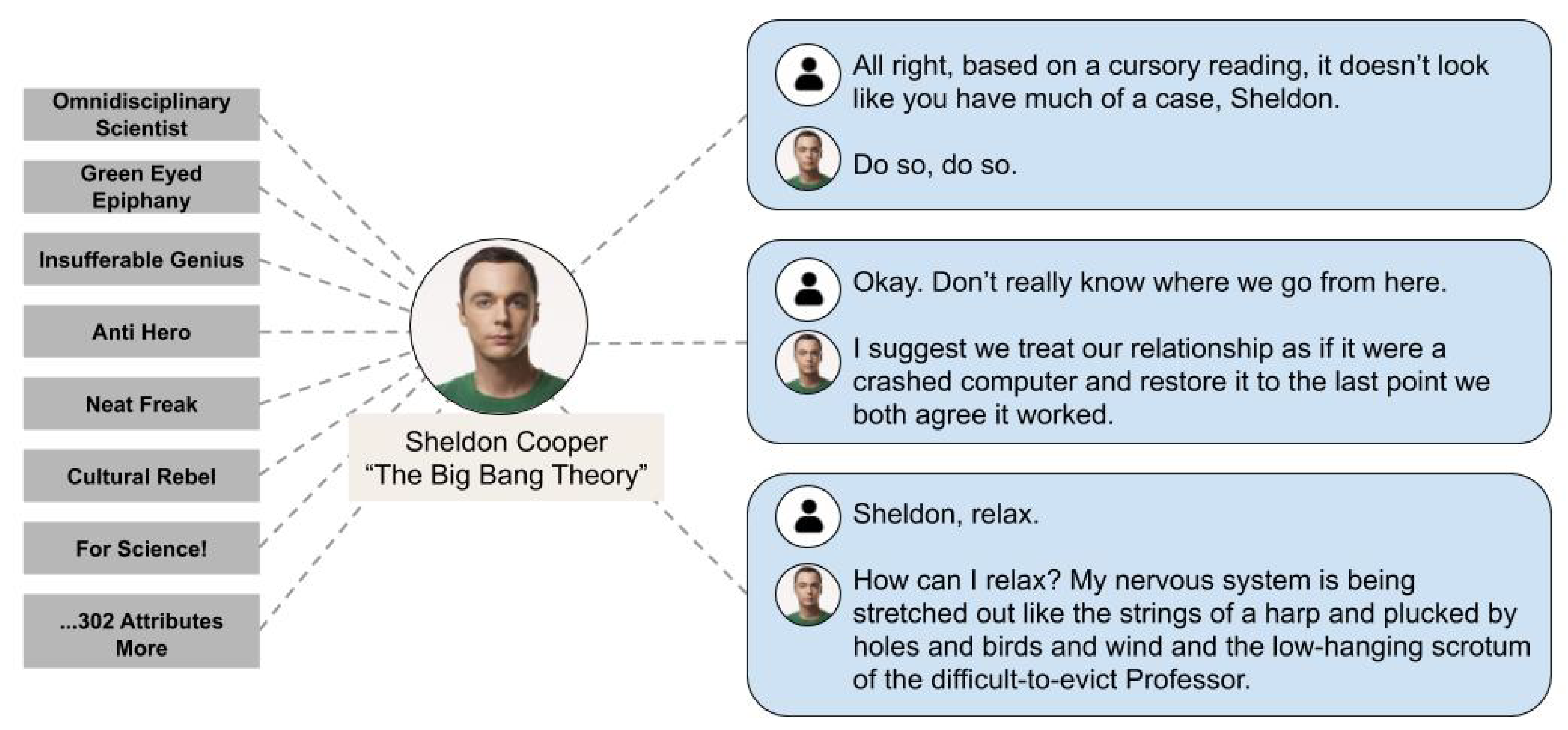
...

[HLA: Cute Bookworm]

Our Dataset: HLA-Chat

- Present a dataset, **HLA-Chat**, with:
  - CHARACTERS
  - ATTRIBUTES (HLAs) of characters
  - Character DIALOGUES

Our Dataset: HLA-Chat



Omnidisciplinary Scientist

Green Eyed Epiphany

Insufferable Genius

Anti Hero

Neat Freak

Cultural Rebel

For Science!

...302 Attributes More

Sheldon Cooper
“The Big Bang Theory”

All right, based on a cursory reading, it doesn’t look like you have much of a case, Sheldon.

Do so, do so.

Okay. Don’t really know where we go from here.

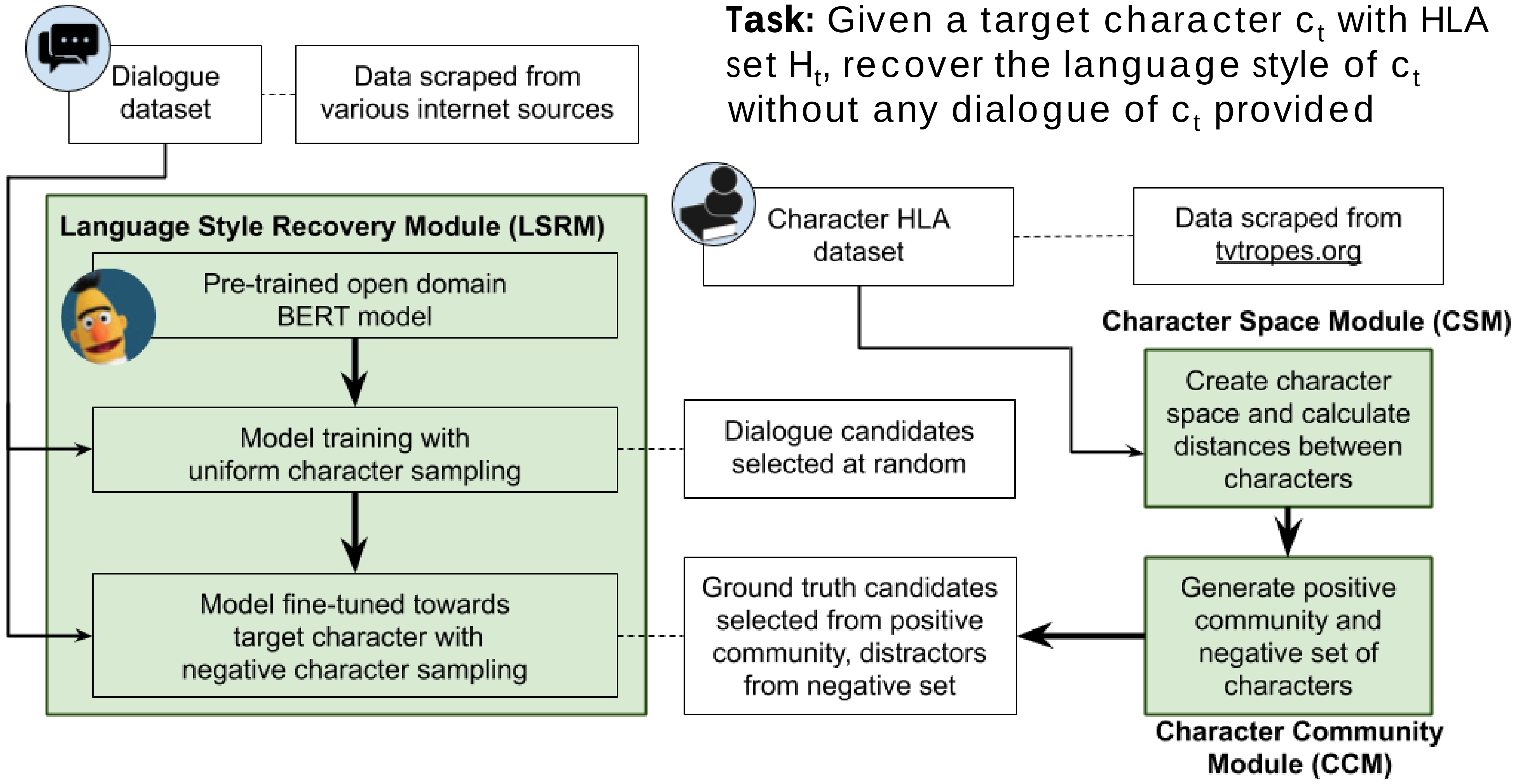
I suggest we treat our relationship as if it were a crashed computer and restore it to the last point we both agree it worked.

Sheldon, relax.

How can I relax? My nervous system is being stretched out like the strings of a harp and plucked by holes and birds and wind and the low-hanging scrotum of the difficult-to-evict Professor.

Our System: ALOHA

Task: Given a target character c_t with HLA set H_t, recover the language style of c_t without any dialogue of c_t provided



Dialogue dataset

Data scraped from various internet sources

Language Style Recovery Module (LSRM)

Pre-trained open domain BERT model

Model training with uniform character sampling

Model fine-tuned towards target character with negative character sampling

Character HLA dataset

Data scraped from tvtropes.org

Dialogue candidates selected at random

Character Space Module (CSM)

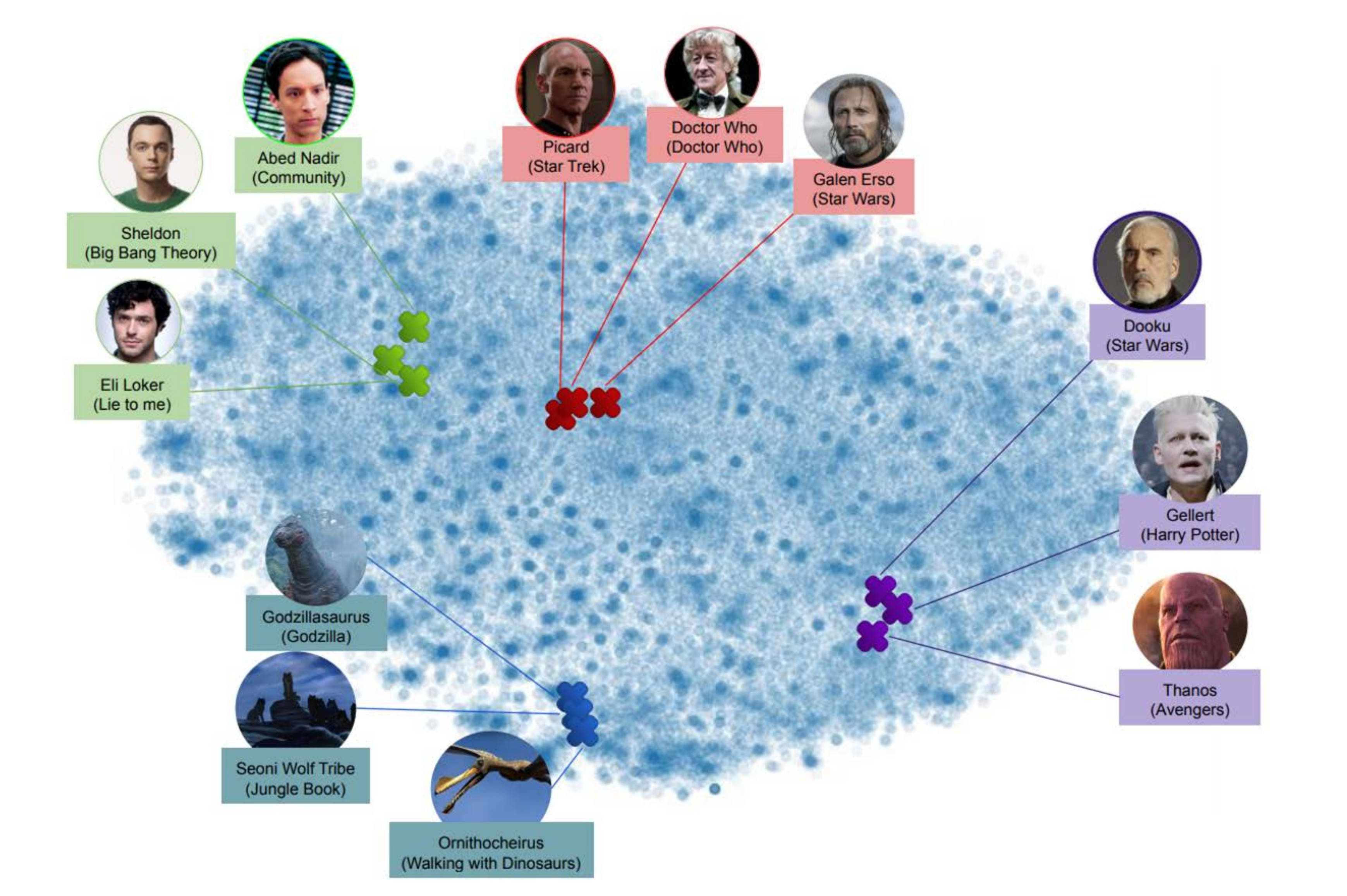
Create character space and calculate distances between characters

Generate positive community and negative set of characters

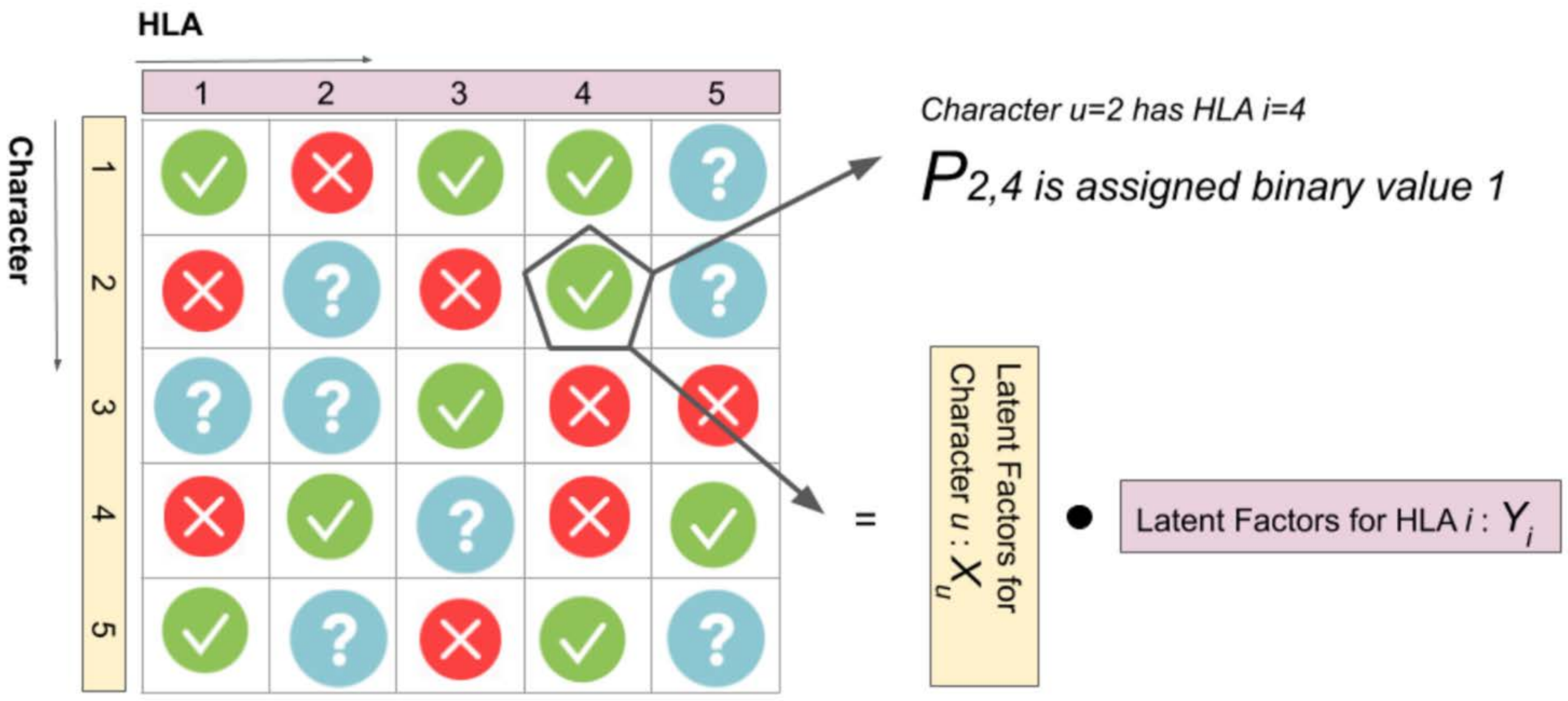
Ground truth candidates selected from positive community, distractors from negative set

Character Community Module (CCM)

Character Space Module (CSM)



CSM: Learning HLAs



Character \ HLA	1	2	3	4	5
1	✓	✗	✓	✓	?
2	✗	?	✗	✓	?
3	?	?	✓	✗	✗
4	✗	✓	?	✗	✓
5	✓	?	✗	✓	?

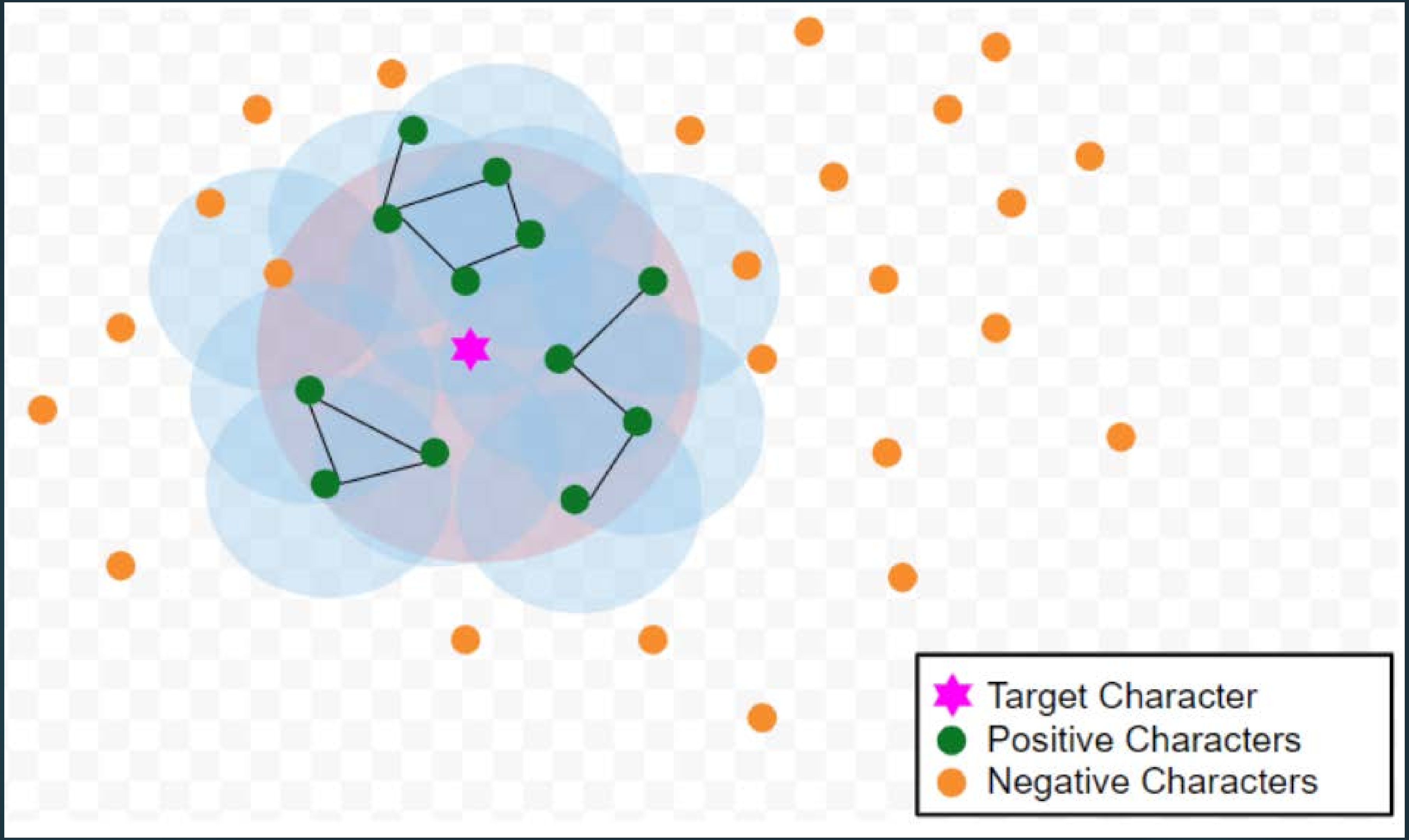
Character u=2 has HLA i=4

$P_{2,4}$ is assigned binary value 1

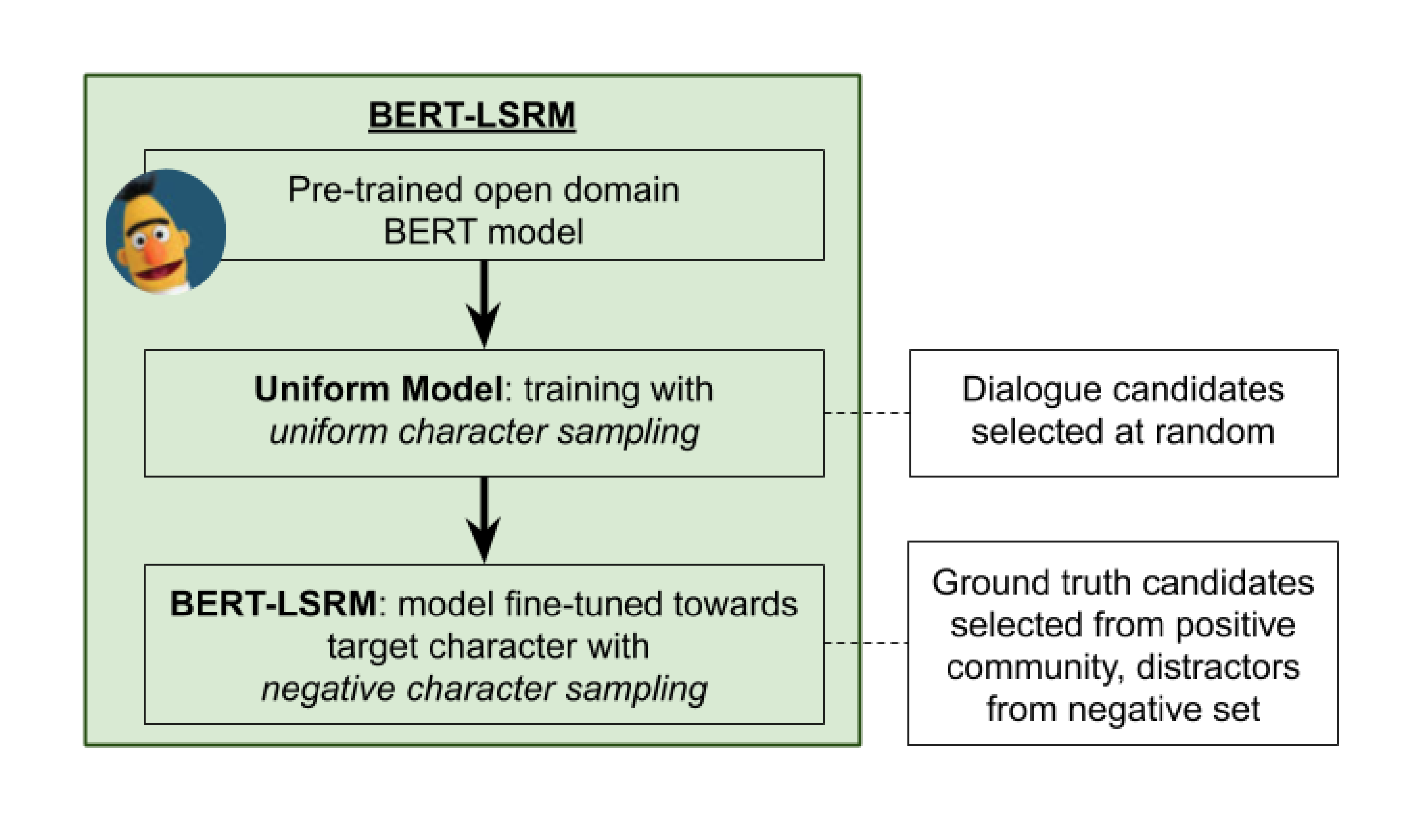
= Latent Factors for Character *u* : X_u ● Latent Factors for HLA *i* : Y_i

$$loss = \sum_u \sum_i (\alpha P_{u,i} - X_u^T Y_i)^2 + \lambda(||X_u||^2 + ||Y_i||^2)$$

Character Community Module (CCM)



Language Style Retrieval Module (LSRM)



BERT-LSRM

Pre-trained open domain BERT model

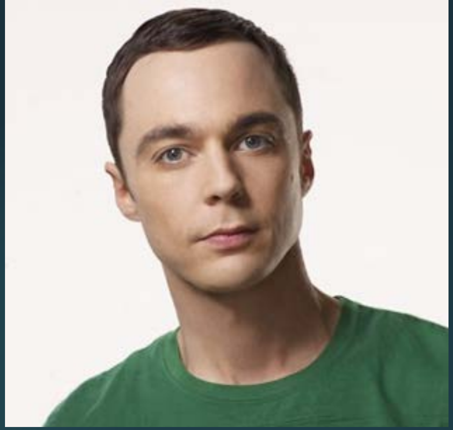
Uniform Model: training with *uniform character sampling*

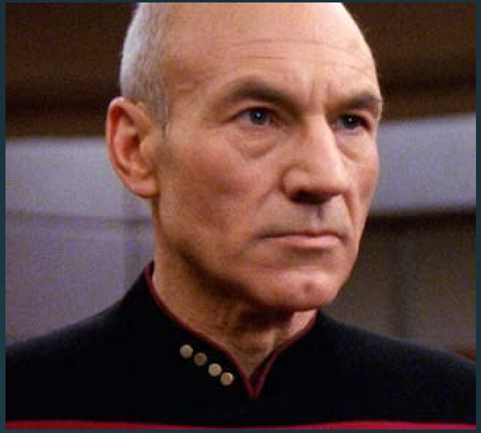
BERT-LSRM: model fine-tuned towards target character with *negative character sampling*

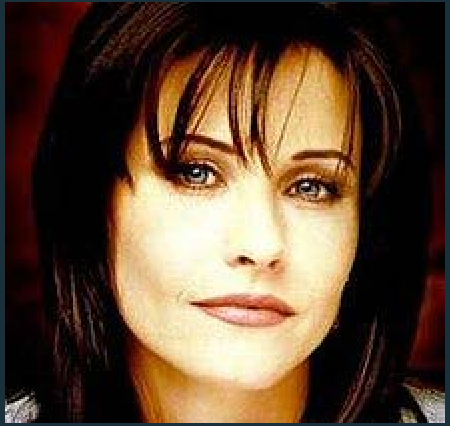
Dialogue candidates selected at random

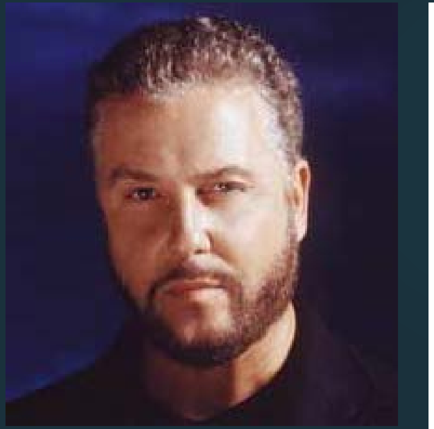
Ground truth candidates selected from positive community, distractors from negative set

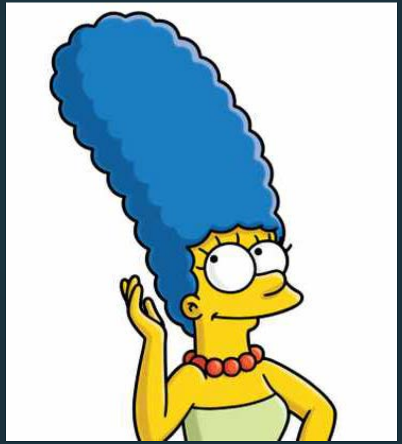
Evaluation: Five Target Characters



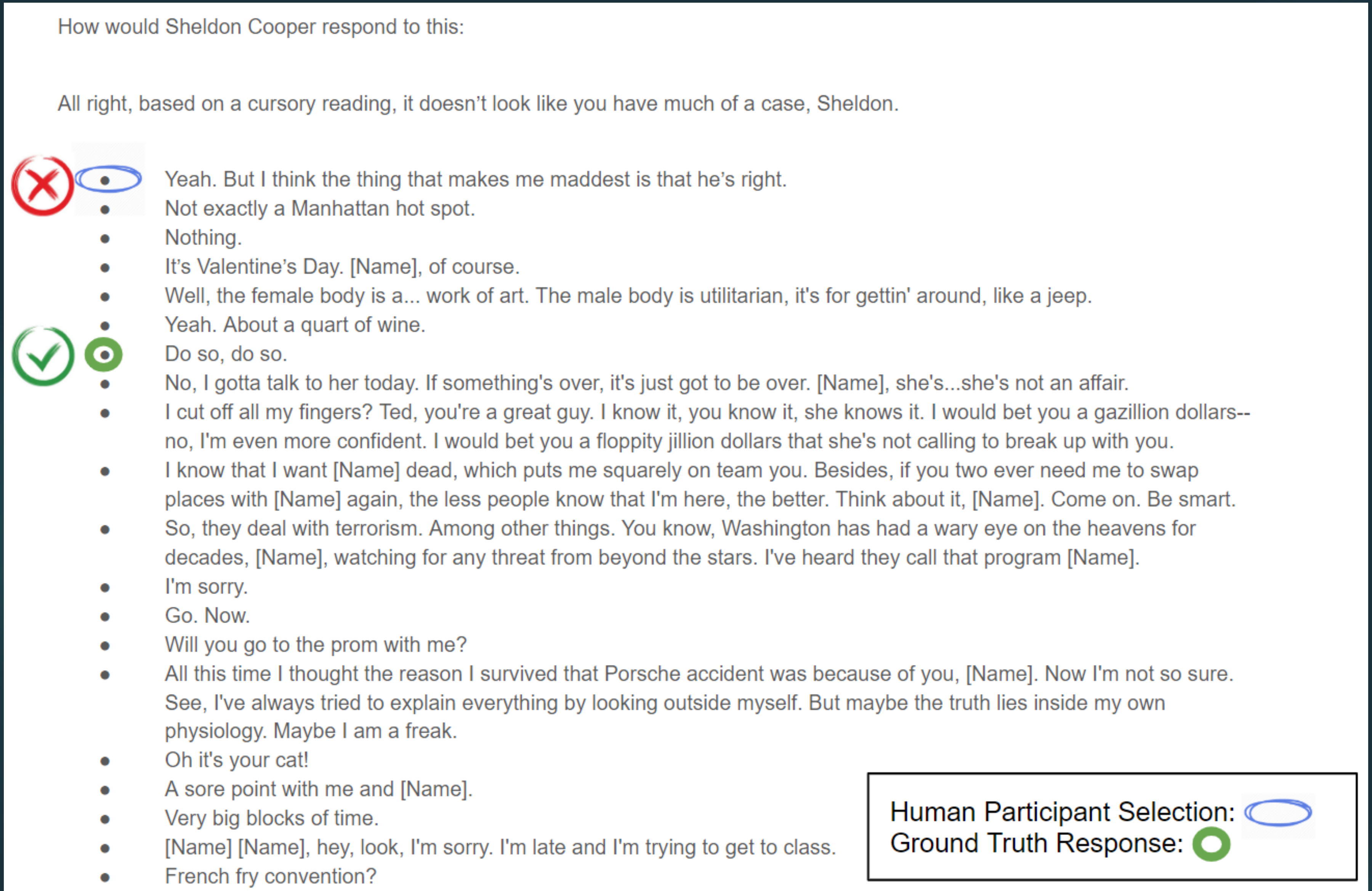








Evaluation Format



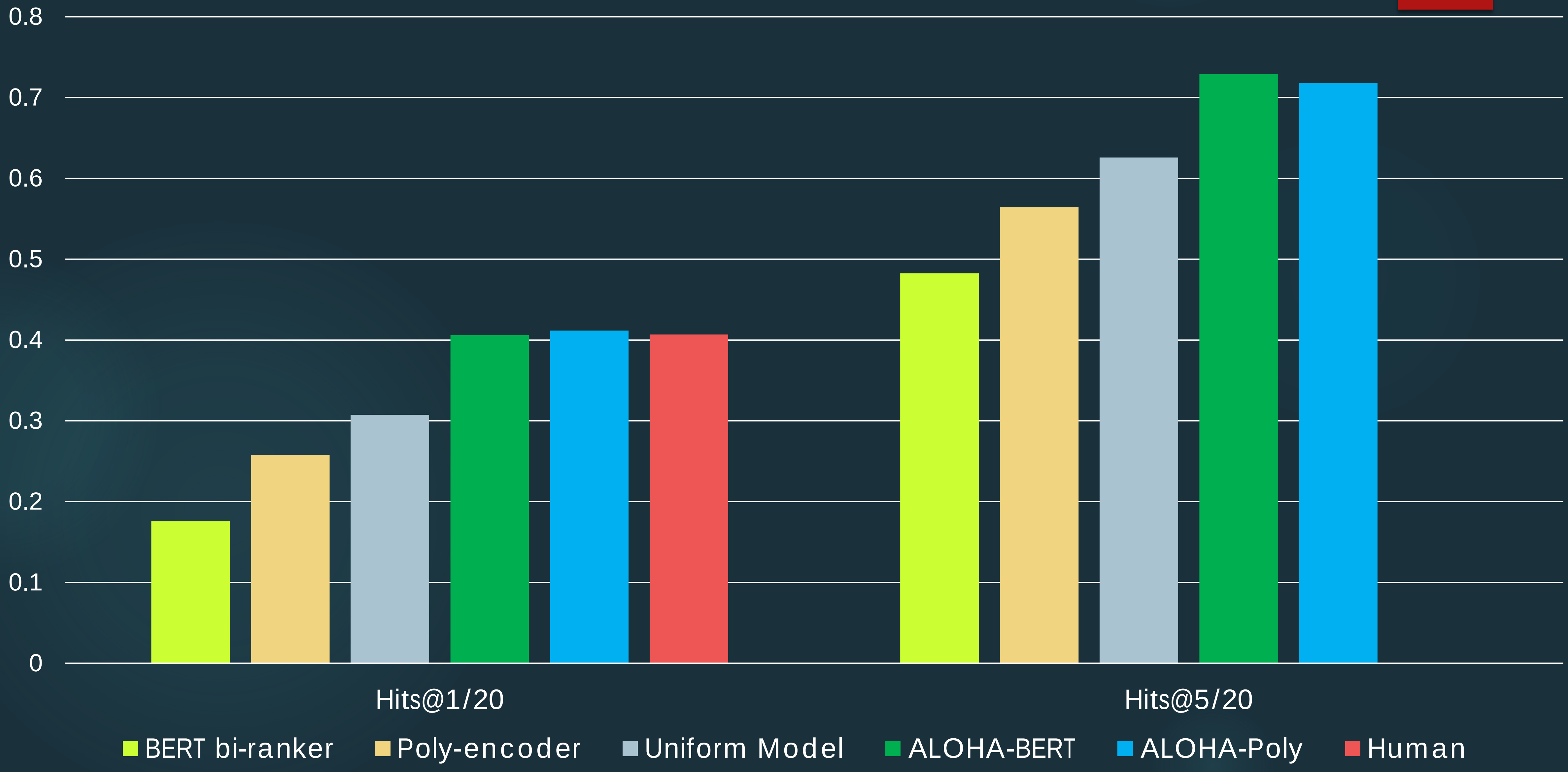
How would Sheldon Cooper respond to this:

All right, based on a cursory reading, it doesn’t look like you have much of a case, Sheldon.

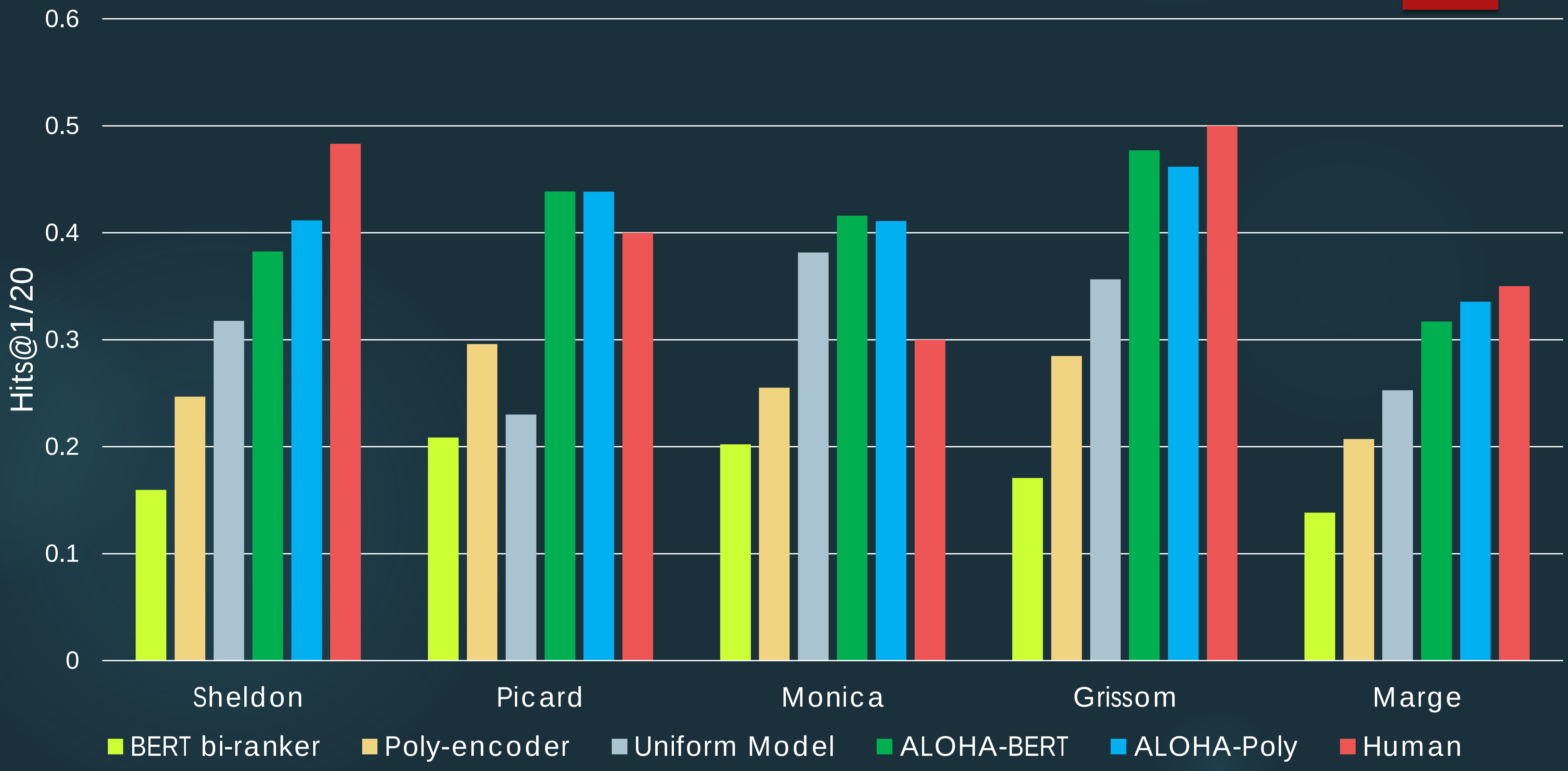
- ✗ ⬭ Yeah. But I think the thing that makes me maddest is that he’s right.
- Not exactly a Manhattan hot spot.
- Nothing.
- It’s Valentine’s Day. [Name], of course.
- Well, the female body is a... work of art. The male body is utilitarian, it's for gettin' around, like a jeep.
- Yeah. About a quart of wine.
- ✓ ◉ Do so, do so.
- No, I gotta talk to her today. If something's over, it's just got to be over. [Name], she's...she's not an affair.
- I cut off all my fingers? Ted, you're a great guy. I know it, you know it, she knows it. I would bet you a gazillion dollars-- no, I'm even more confident. I would bet you a floppity jillion dollars that she's not calling to break up with you.
- I know that I want [Name] dead, which puts me squarely on team you. Besides, if you two ever need me to swap places with [Name] again, the less people know that I'm here, the better. Think about it, [Name]. Come on. Be smart.
- So, they deal with terrorism. Among other things. You know, Washington has had a wary eye on the heavens for decades, [Name], watching for any threat from beyond the stars. I've heard they call that program [Name].
- I'm sorry.
- Go. Now.
- Will you go to the prom with me?
- All this time I thought the reason I survived that Porsche accident was because of you, [Name]. Now I'm not so sure. See, I've always tried to explain everything by looking outside myself. But maybe the truth lies inside my own physiology. Maybe I am a freak.
- Oh it's your cat!
- A sore point with me and [Name].
- Very big blocks of time.
- [Name] [Name], hey, look, I'm sorry. I'm late and I'm trying to get to class.
- French fry convention?

Human Participant Selection: ⬭
Ground Truth Response: ◉

Average Evaluation Results (Hits@1/20 and 5/20)



Average Hits@1/20 by Evaluation Character



Future Work

- Model conversations with a multi-turn response approach
- Model the dialogue counterpart
- HLA-aligned generative models
- Determining HLAs from text
- Larger and more extensive human evaluation

Semantic Text Exchange (STE)

- Task: correct the semantic content of text (e.g. personalized dialogue agent responses) using the original text as a template
- Goal: correct semantics while preserving sentiment and fluency
- Example:

User input:	***What is the weather like today?***
Original output text (negative emotion):	*It is **sunny** outside. I know, it sucks! But you should **wear sunscreen** even if it's **sticky**.*
Replacement entity:	*Rainy*
Desired output text (negative emotion):	*It is **rainy** outside. I know, it sucks! But you should **bring an umbrella** even if it's **cumbersome**.*

- SMERTI: Similarity Masking, Entity Replacement, Text Infilling

Feng, Li, and Hoey. ***Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange***. Proceedings of EMNLP 2019. aclweb.org/anthology/D19-1272/

Summary

- **Major Challenge** - Giving chatbots and virtual assistants the ability to imitate human emotions and personality
- **Human-Level Attributes (HLAs)** - Train dialogue agents with tropes: aspects of fictional characters' identity and profile
  - HLAs accurately model human impressions of characters
- **HLA-Chat** - Dataset of characters with their HLAs and dialogue
- **Artificial Learning of Human Attributes** (**ALOHA)** - Three-stage system to retrieve character/personality specific language styles
  - ALOHA effectively uses HLAs to retrieve character specific language styles
  - ALOHA shown to be stable and robust

Thanks for Listening!

https://github.com/newpro/aloha-chatbot

https://arxiv.org/abs/1910.08293

- Steven Y. Feng: stevenyfeng@gmail.com

  Website: https://styfeng.github.io/

  Twitter: @stevenyfeng

- Aaron W. Li: w89li@uwaterloo.ca

- Veronica Jiang: r4jiang@uwaterloo.ca

- Julia Sprague: jsprague@edu.uwaterloo.ca

- Wei Zhou: wei.zhou1@huawei.com

- Jesse Hoey: jhoey@cs.uwaterloo.ca

  Website: https://cs.uwaterloo.ca/~jhoey/

  Twitter: @drjessehoey

Input: ***What do you think of my AAAI paper?***

Sheldon (Reviewer 1): *Very good. Author of one of the greatest horror stories ever written: Turn of the Screw. And I'm looking for one.*
Picard (Reviewer 2): *Class M in all respects. Quite Earth-like.*
Marge (Reviewer 3): *I think it's awesome. I love it.*

Input: ***Let us go to a party!***

Sheldon: *No chance, hot pants! I'd rather die and kill all of you than party for one more millisecond!*
Picard: *Whoa, whoa, whoa! I enjoy partying as much as the next fellow, but this is Oktoberfest. Good day, sir!*
Marge: *I think you cannot have a party. Your mother left me in charge.*

Appendices

Mathematical Analysis

- Pearson correlation values:
  - Human vs. Uniform: 0.047
  - Human vs. ALOHA: 0.4149 and 0.5468

- Paired t-test p-value results (of Hits@1/20 scores):
  - ALOHA-BERT vs. BERT Bi-ranker: 0.0004
  - ALOHA-Poly vs. Poly-encoder: <0.0001
  - ALOHA-BERT vs. Uniform: 0.0329
  - ALOHA-Poly vs. Uniform: 0.0234