Public Self-consciousness for Endowing Dialogue Agents with Consistent Persona

ICLR 2020 *BAICS workshop* (Oral)

Hyunwoo Kim
Byeongchang Kim
Gunhee Kim

VISION & LEARNING LAB
SEOUL NATIONAL UNIVERSITY

The Consistency Problem

in Dialogue Agents

Human: What is your job?

***Bot**: I'm a programmer.*

Human: What do you do?

***Bot**: I'm a lawyer.*

Human: ???

Previous works
tackling the Consistency Problem

Embeddings

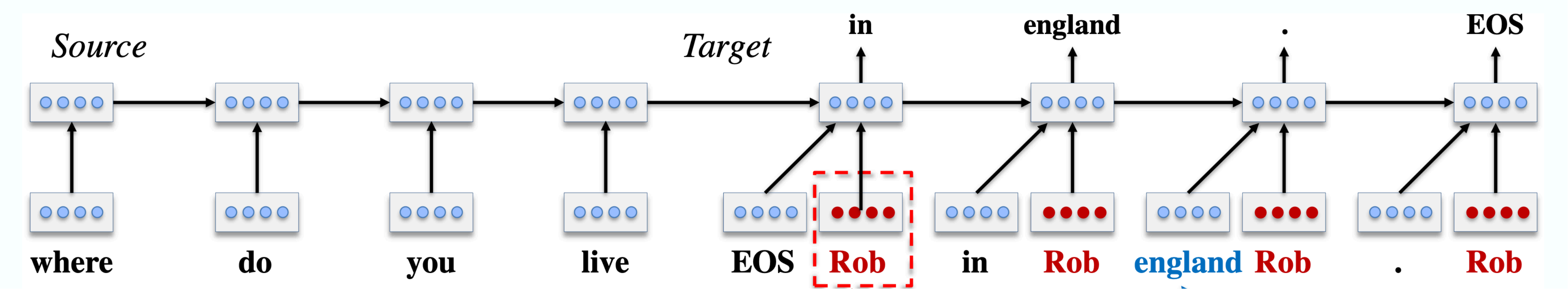
Benchmark Datasets

Natural Language Inference (NLI)

Previous Works:

Input persona embeddings to the model

- **Feed a persona embedding** to the decoder along with the target utterance



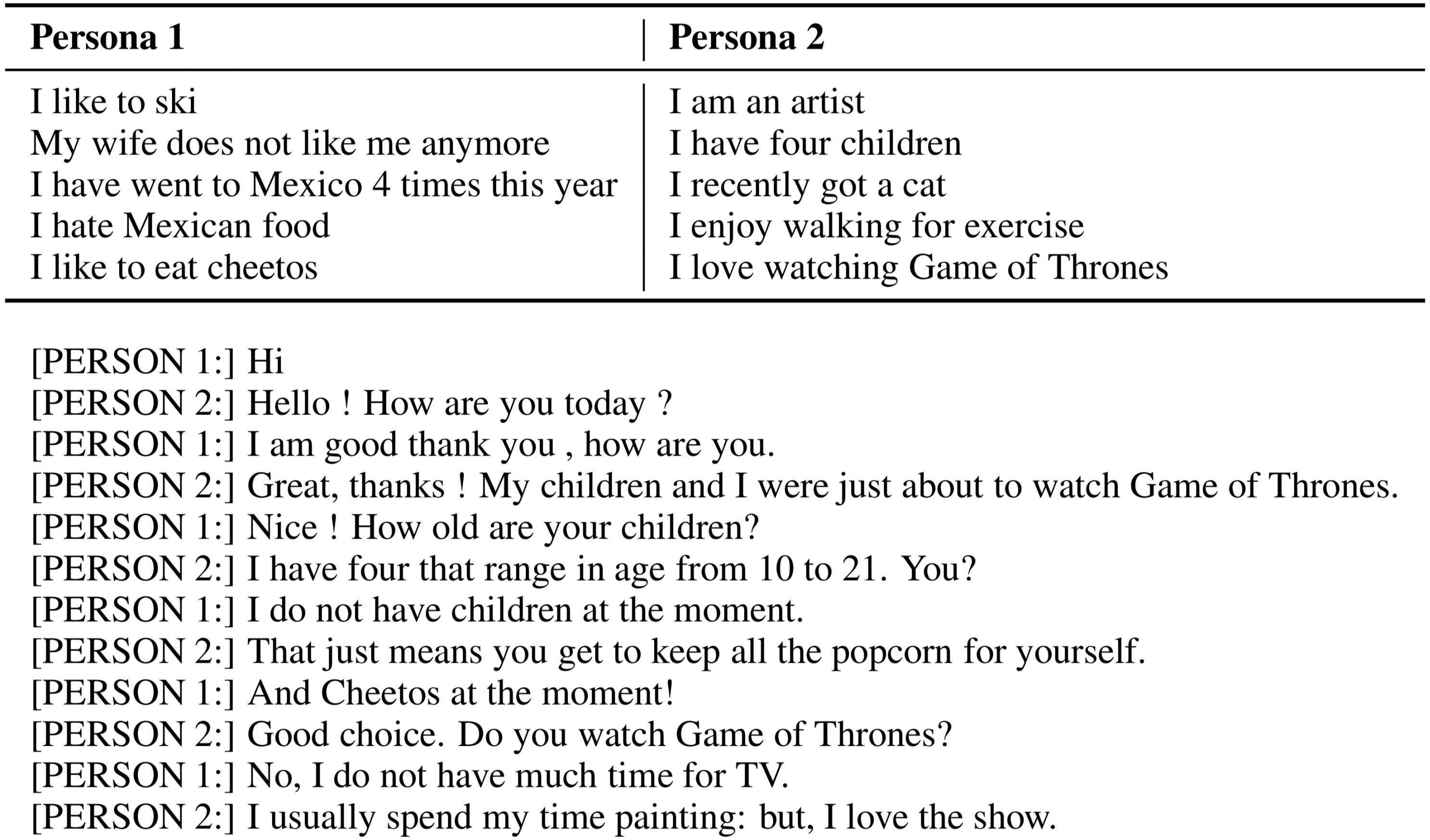
Dolan et al. 2016. A persona-based neural conversational model. *ACL*

Previous Works:

Benchmark dataset which persona sentences are given to the model

- **the *PersonaChat* dataset**

A dialogue dataset involving
two interlocutors
getting to know each other
while playing the given persona



Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Zhang et al. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *ACL*

Previous Works:

Exploit Natural Language Inference (NLI) annotations

Given a "premise",
the task of determining whether a "hypothesis" is

- True (Entailment)
- False (Contradiction)
- Undetermined (Neutral)

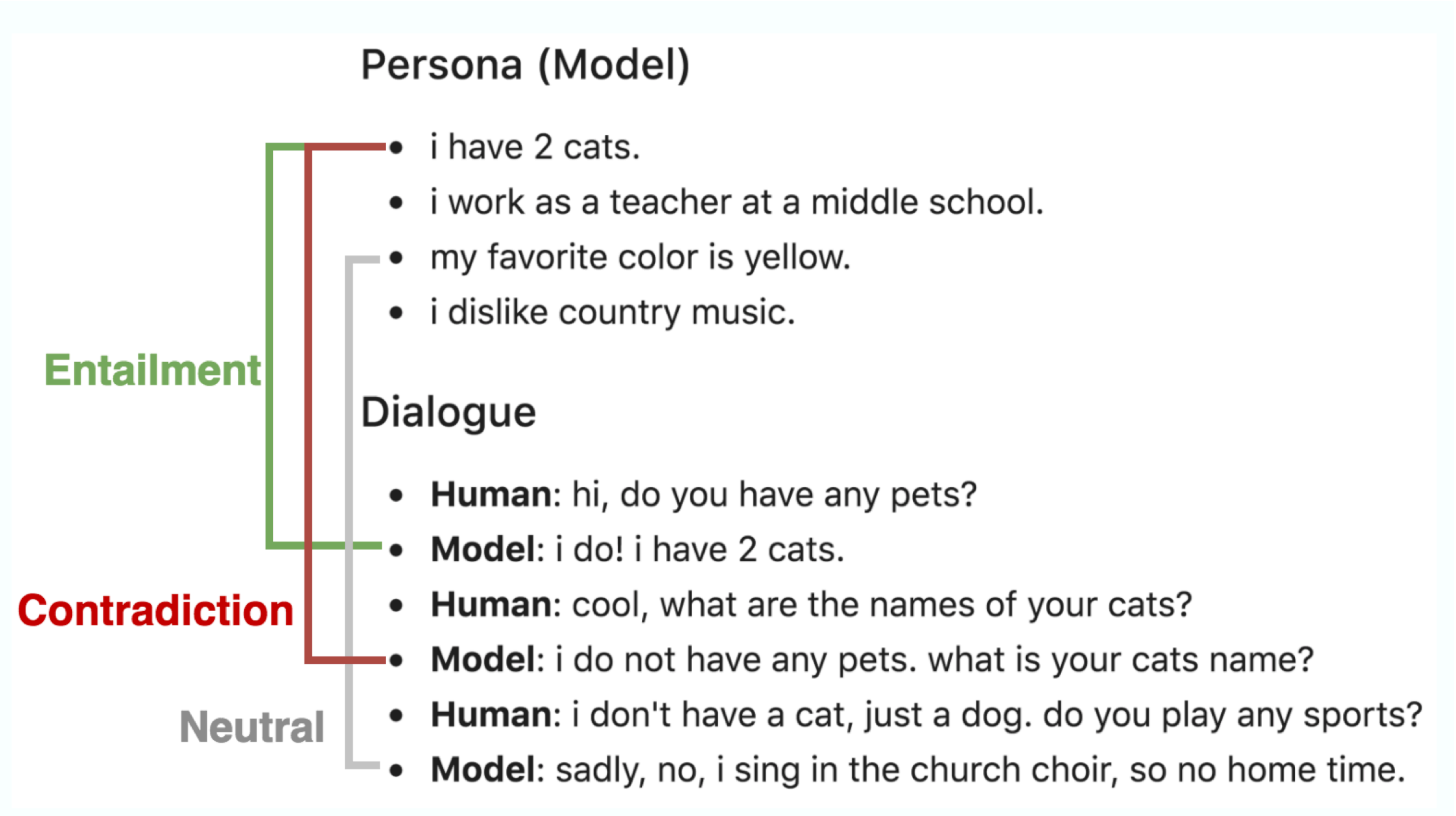
Premise: I love to go for a drive with my new car.

- Hypothesis: Recently, I finally bought a car! **[Entailment]**
- Hypothesis: I do not have a car. **[Contradiction]**
- Hypothesis: Milk shake is my favorite dessert. **[Neutral]**

Welleck et al. 2019. Dialogue Natural Language Inference. *ACL*

Previous Works: use NLI

1. collect **additional NLI annotations**



Persona (Model)

- i have 2 cats.
- i work as a teacher at a middle school.
- my favorite color is yellow.
- i dislike country music.

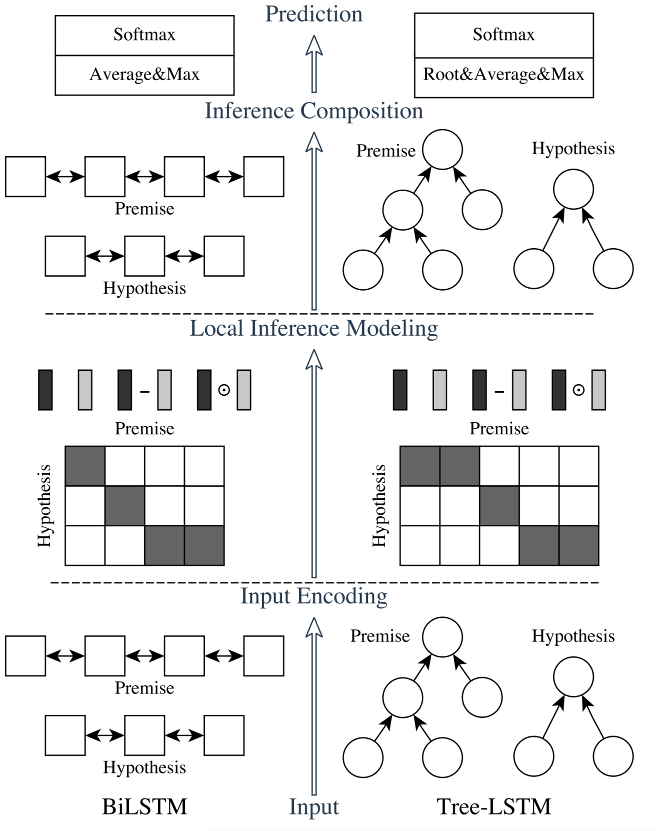
Dialogue

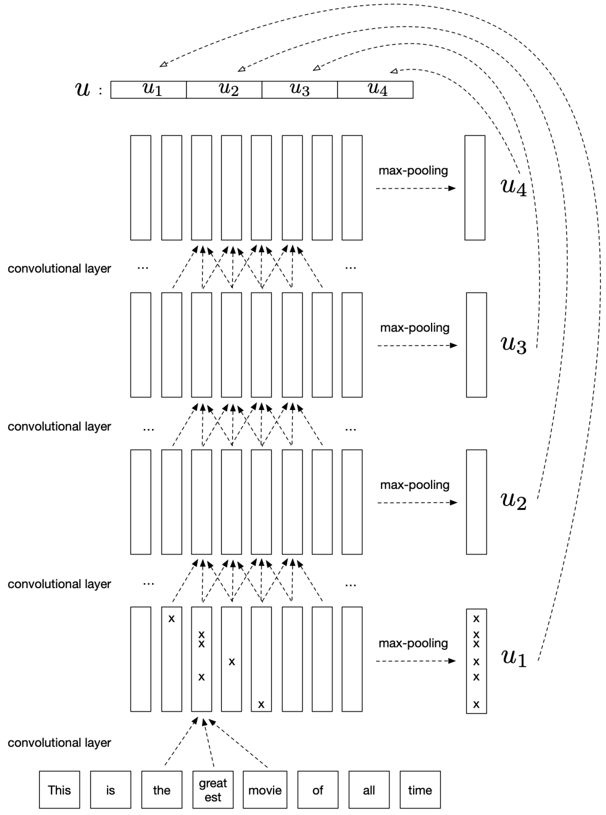
- **Human**: hi, do you have any pets?
- **Model**: i do! i have 2 cats.
- **Human**: cool, what are the names of your cats?
- **Model**: i do not have any pets. what is your cats name?
- **Human**: i don't have a cat, just a dog. do you play any sports?
- **Model**: sadly, no, i sing in the church choir, so no home time.

Welleck et al. 2019. Dialogue Natural Language Inference. *ACL*

Previous Works: use NLI

2. train **external NLI model** on the annotation



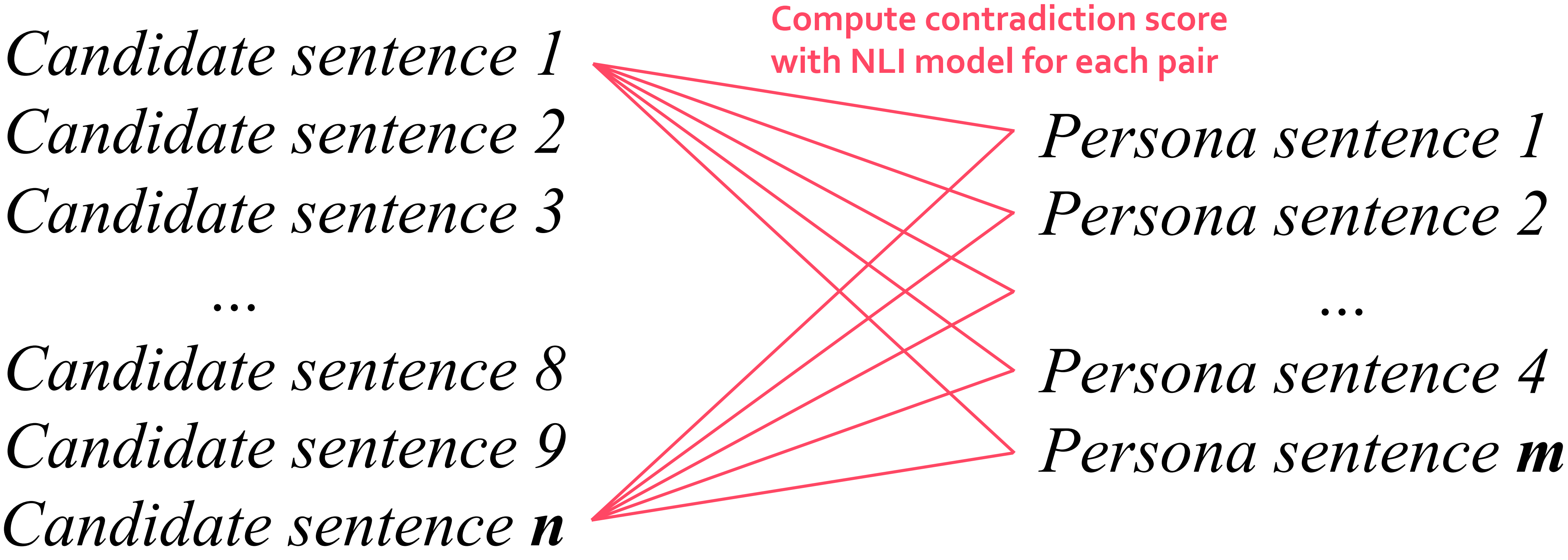


Chen et al. 2017. Enhanced LSTM for Natural Language Inference. *EMNLP* (left)

Conneau et al. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *ACL* (right)

Previous Works: use NLI

3. compute **pair-wise contradiction scores** on **every** candidate sentences of the dialogue agent and persona sentences to *re-weight* contradicting candidates



Compute contradiction score with NLI model for each pair

Candidate sentence 1
Candidate sentence 2
Candidate sentence 3
…
Candidate sentence 8
Candidate sentence 9
Candidate sentence ***n***

Persona sentence 1
Persona sentence 2
…
Persona sentence 4
Persona sentence ***m***

Welleck et al. 2019. Dialogue Natural Language Inference. *ACL*

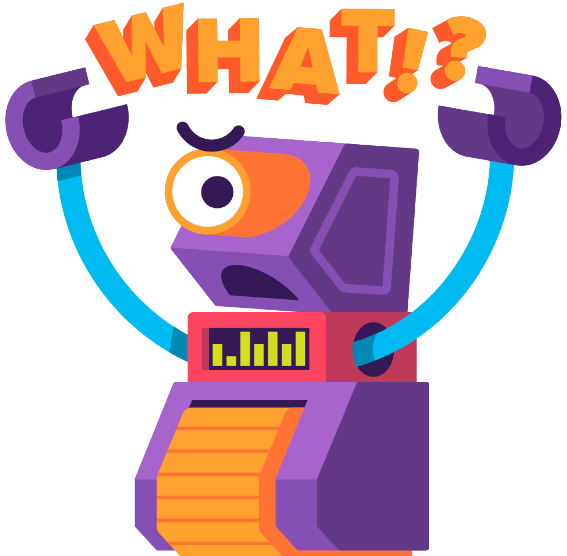
Song et al. 2019. Generating Persona Consistent Dialogues by Exploiting Natural Language Inference. *arXiv*

Previous Works: use NLI

Limitations

1. Require **NLI annotations** *on the target dataset*
2. Require training **external NLI model** on the annotations
3. NLI model computes **pair-wise contradiction score** *for every* persona sentences and candidate sentences

⇨ **Demanding & Inscalable**



Our question:

How do humans maintain consistency?

We do not ask others
whether we are consistent or not

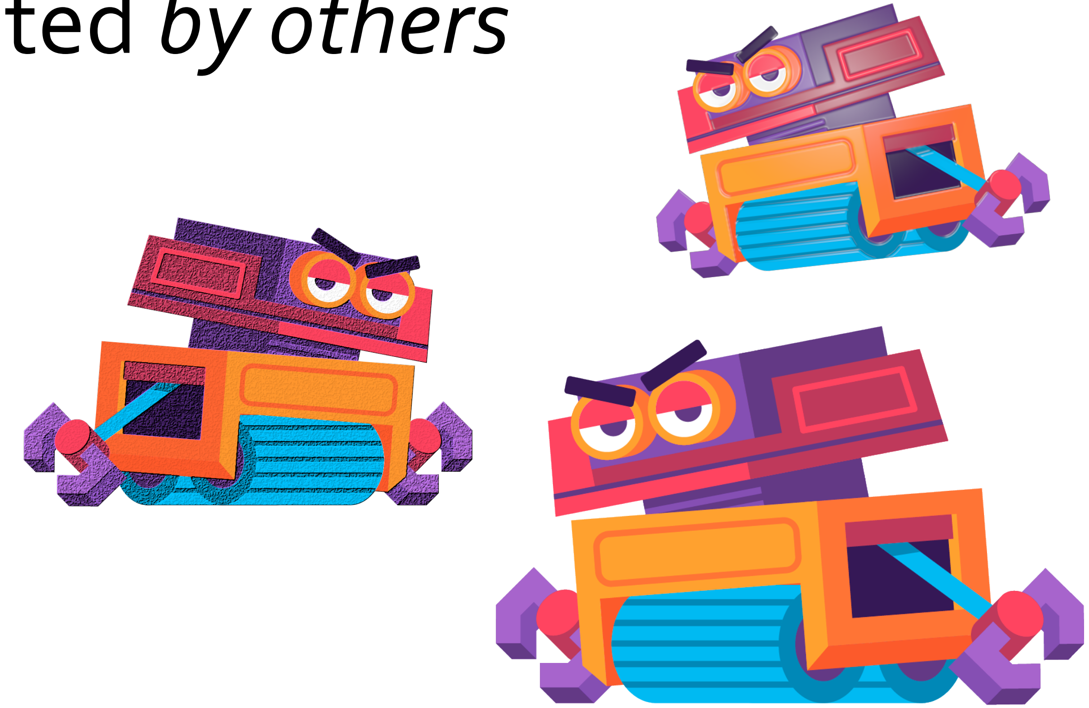
We ask ourselves.

We ask ourselves.

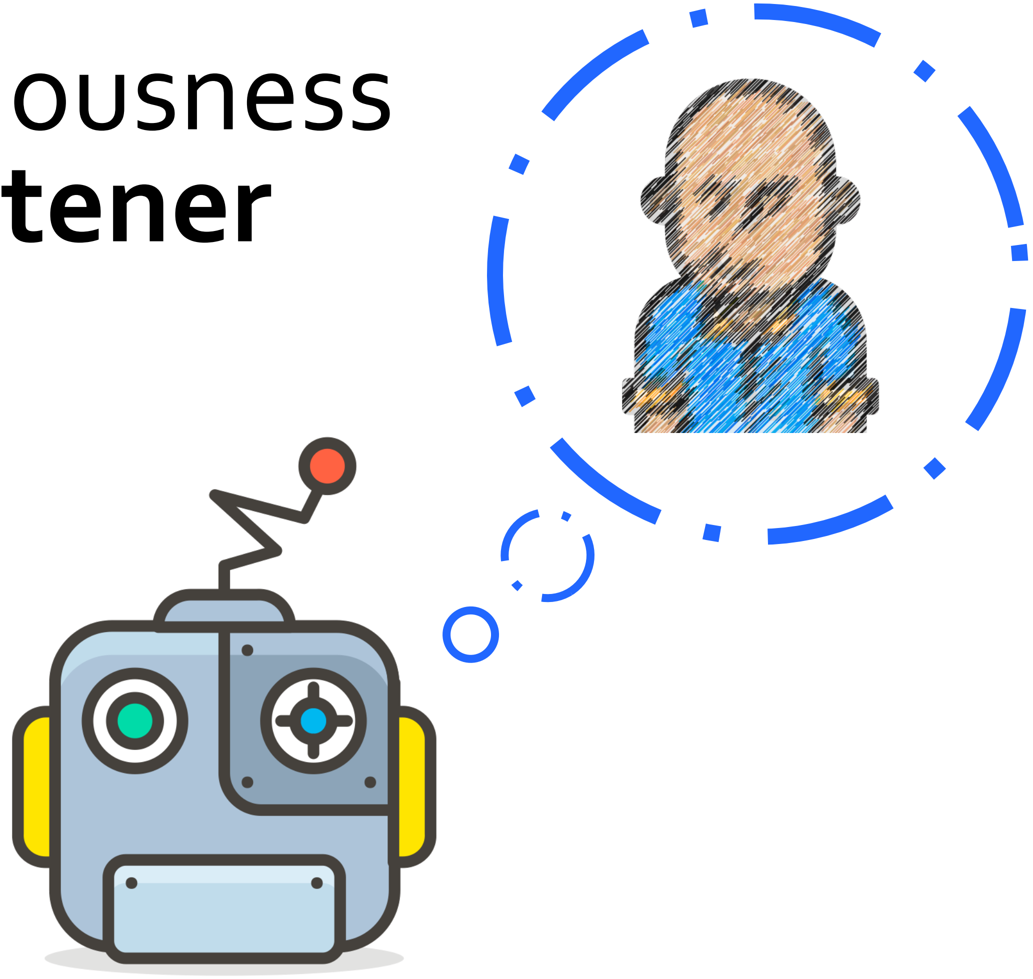
by predicting
how we will be perceived by others

Public Self-Consciousness

The *awareness of the self* as a social object that can be observed and evaluated *by others*



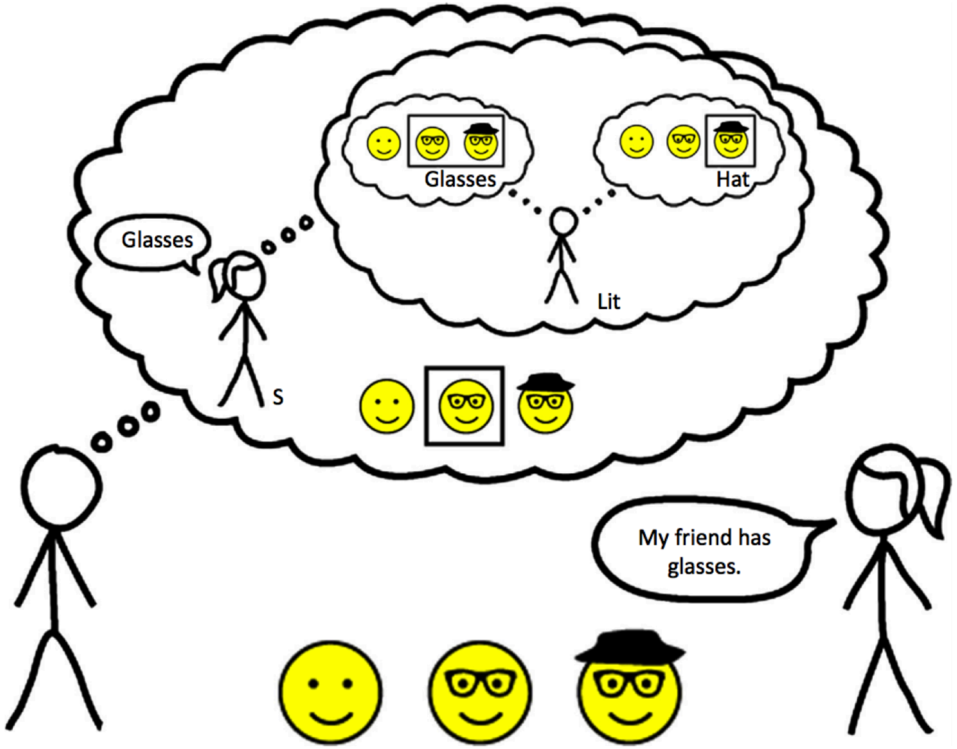
We model the self-consciousness through an **imaginary listener**



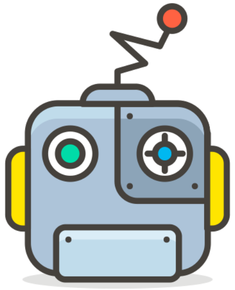
Modeling a Listener:

The Bayesian Rational Speech Acts framework

Treats language use as a recursive process
where probabilistic speaker and listener reason about each other
in Bayesian fashion



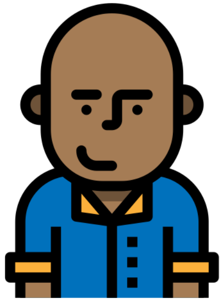
Frank and Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*



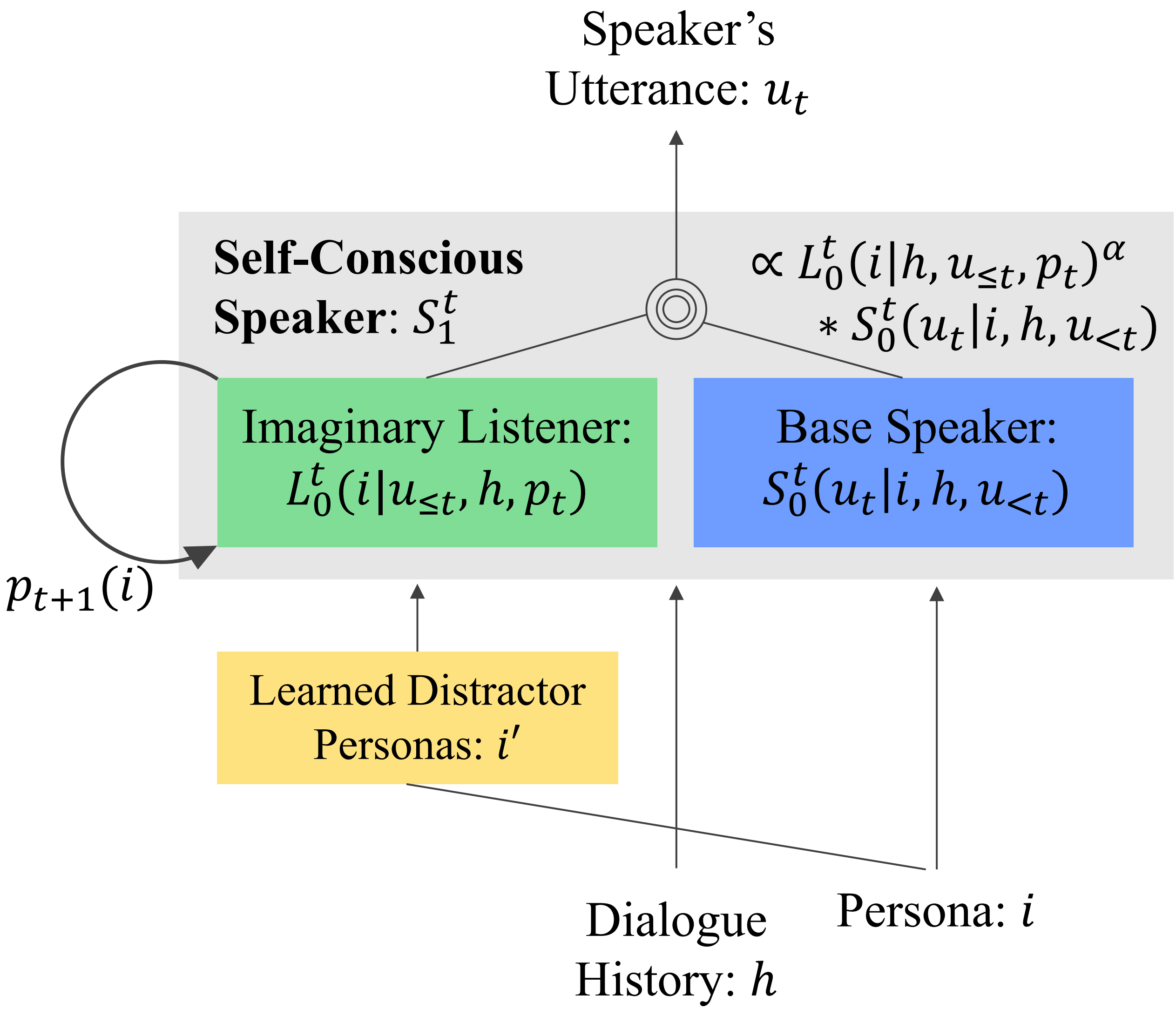
Our approach:

A self-conscious agent

thinking about how it will be perceived



The Self-Conscious Speaker S_1


Speaker's Utterance: u_t
Self-Conscious Speaker: S_1^t
$\propto L_0^t(i|h, u_{\leq t}, p_t)^{\alpha}$
$* S_0^t(u_t|i, h, u_{<t})$
Imaginary Listener: $L_0^t(i|u_{\leq t}, h, p_t)$
Base Speaker: $S_0^t(u_t|i, h, u_{<t})$
$p_{t+1}(i)$
Learned Distractor Personas: i'
Dialogue History: h
Persona: i

Task Setting:

🤖’s Persona (Speaker 1’s Persona)

I live in Florida and have a dog.
I am going to college next year.
I enjoy going outside and playing with my friends.
I love Disney movies and animations.

$i: \mathit{given\ persona}$

[Speaker 2] Hello, how are you today?
[Speaker 1] Great! Just watching my favorite TV show. You?
[Speaker 2] Cool! What do you like to do when COVID’s over?

$h: \mathit{dialogue\ history}$

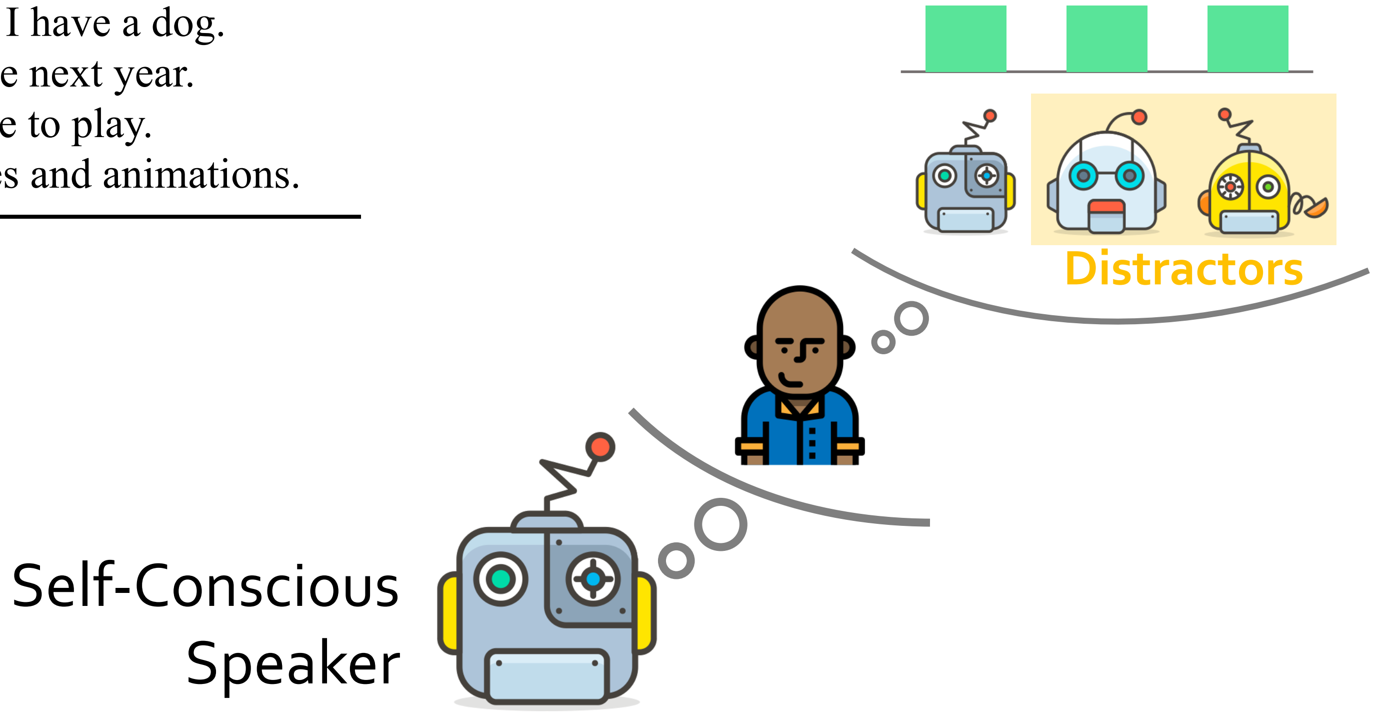
🤖 [Model’s generation]: $u_1, u_2, u_3, \ldots, u_{t-1}, u_t$

$u: \mathit{utterance}\ (t\ \mathit{tokens})$

Intuitive Explanation of the Self-Conscious Speaker S_1

🤖’s Persona

I live in Florida and I have a dog.
I am going to college next year.
I enjoy going outside to play.
I love Disney movies and animations.



‘Will I sound like me?’

‘I want to be identified as my persona, not some other different persona.’

🤖’s Persona

I like reading books.
I raise two cats.
My girlfriend is a developer.
I like to eat pepperoni pizza.

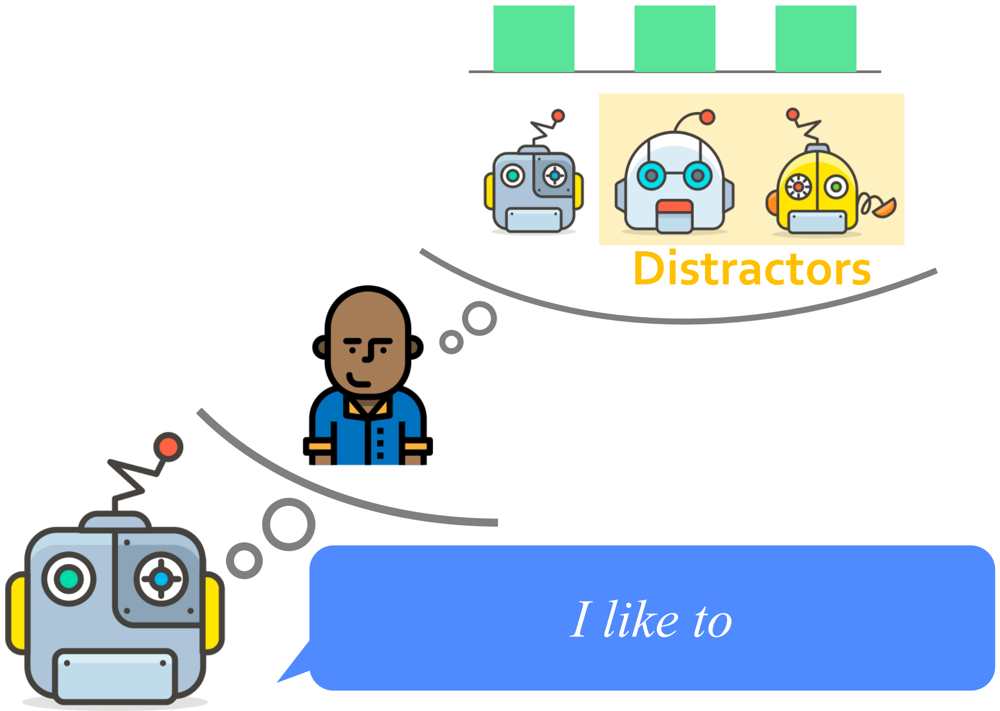
🤖’s Persona

I live in a big city
I work at the gym as a trainer.
I have two dogs.
I like to watch extreme sports.

Intuitive Explanation of the Self-Conscious Speaker S_1

's Persona

I live in Florida and I have a dog.
I am going to college next year.
I enjoy going outside to play.
I love Disney movies and animations.



Self-Conscious Speaker

'Will I sound like me?'

'I want to be identified as my persona, not some other different persona.'

's Persona

I like reading books.
I raise two cats.
My girlfriend is a developer.
I like to eat pepperoni pizza.

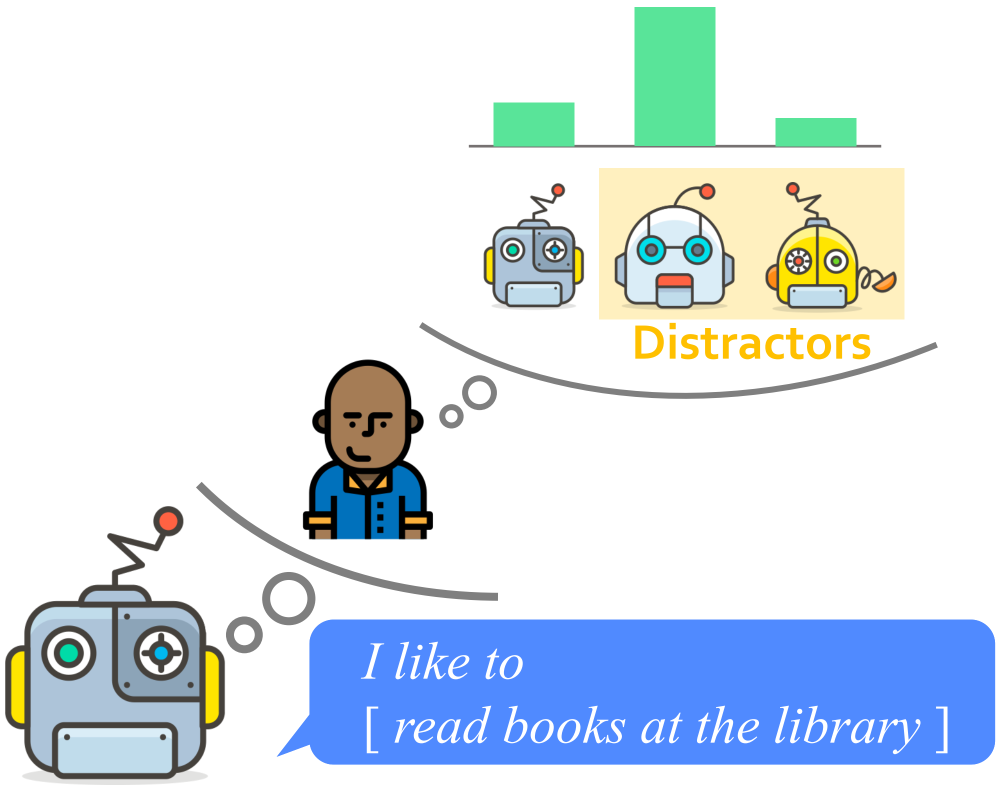
's Persona

I live in a big city
I work at the gym as a trainer.
I have two dogs.
I like to watch extreme sports.

Intuitive Explanation of the Self-Conscious Speaker S_1

’s Persona

I live in Florida and I have a dog.
I am going to college next year.
I enjoy going outside to play.
I love Disney movies and animations.



Self-Conscious Speaker

I like to
[read books at the library]

‘Will I sound like me?’

‘I want to be identified as my persona, not some other different persona.’

’s Persona

I like reading books.
I raise two cats.
My girlfriend is a developer.
I like to eat pepperoni pizza.

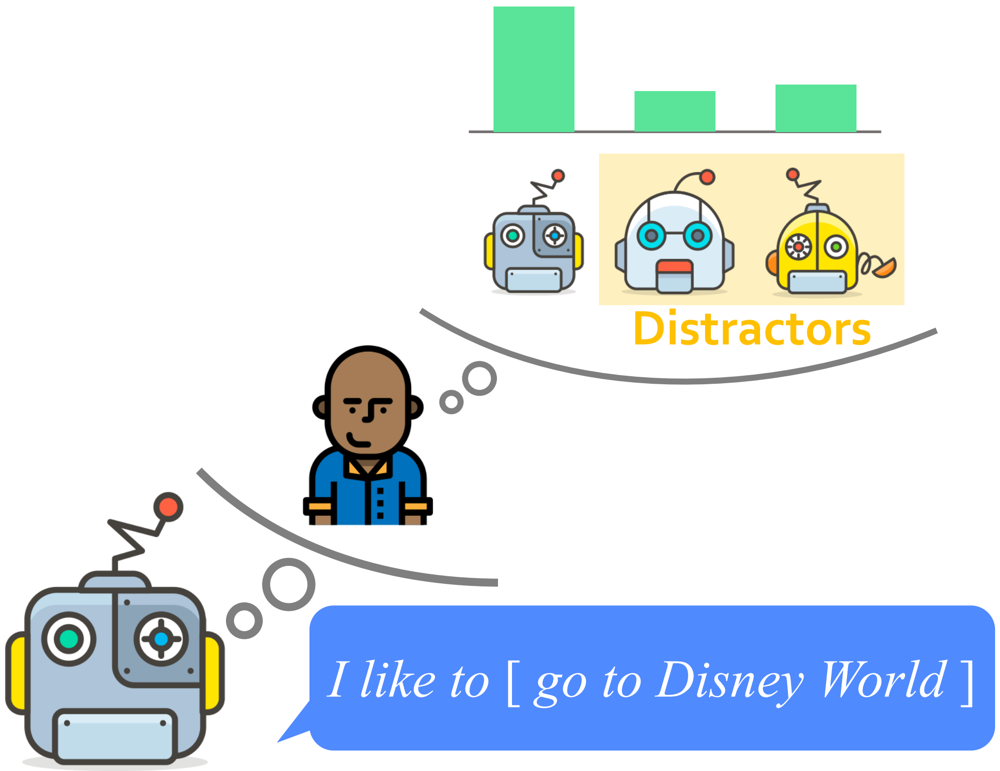
’s Persona

I live in a big city
I work at the gym as a trainer.
I have two dogs.
I like to watch extreme sports.

Intuitive Explanation of the Self-Conscious Speaker S_1

’s Persona

I live in Florida and I have a dog.
I am going to college next year.
I enjoy going outside to play.
I love Disney movies and animations.



I like to [go to Disney World]

Self-Conscious Speaker

‘Will I sound like me?’

‘I want to be identified as my persona, not some other different persona.’

’s Persona

I like reading books.
I raise two cats.
My girlfriend is a developer.
I like to eat pepperoni pizza.

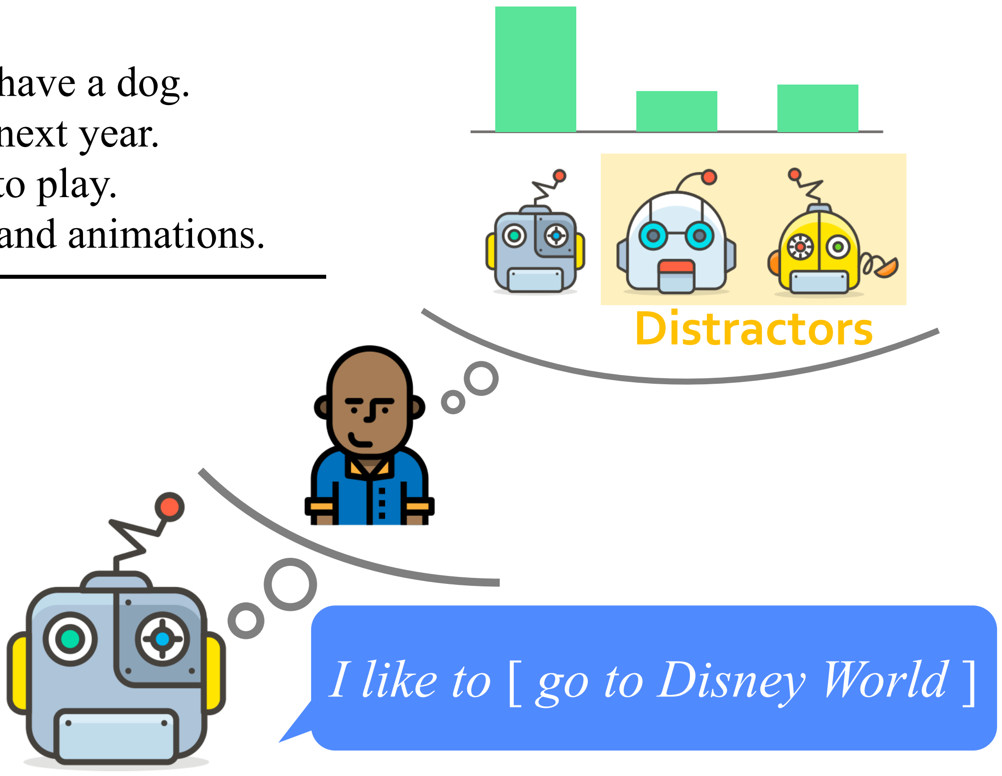
’s Persona

I live in a big city
I work at the gym as a trainer.
I have two parrots.
I like to watch extreme sports.

Intuitive Explanation of the Self-Conscious Speaker S_1

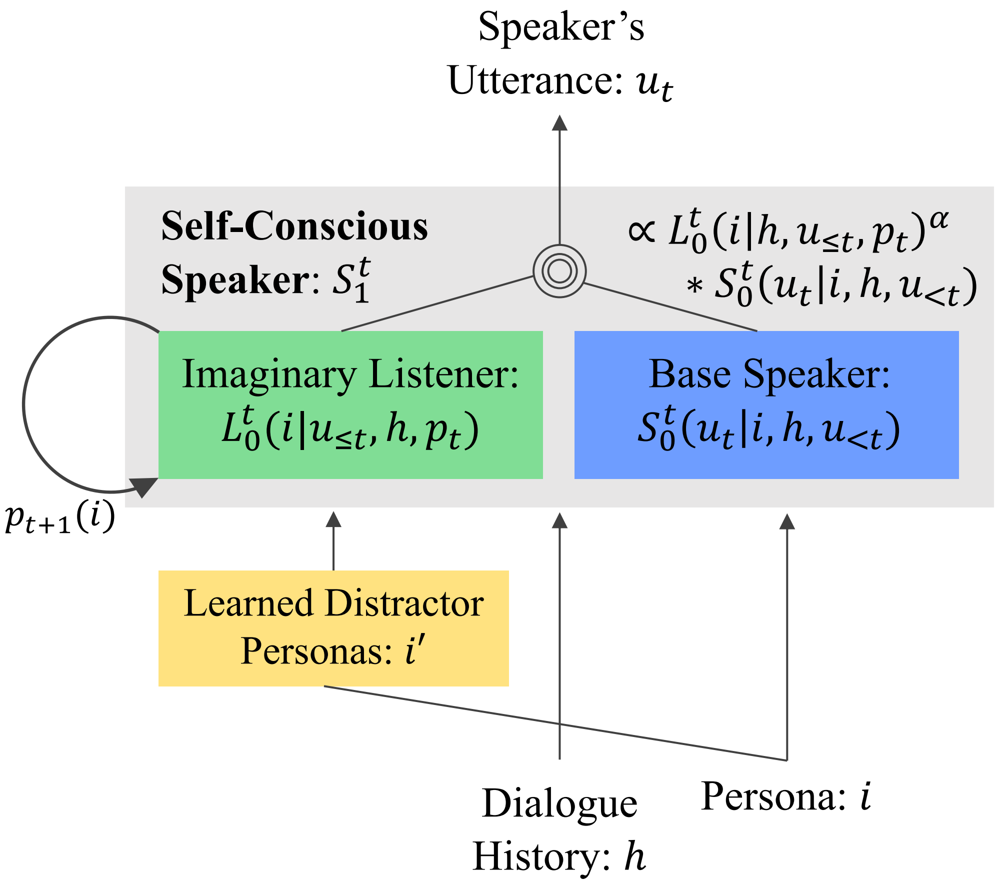
’s Persona

I live in Florida and I have a dog.
I am going to college next year.
I enjoy going outside to play.
I love Disney movies and animations.



‘Will I sound like me?’

‘I want to be identified as my persona, not some other different persona.’

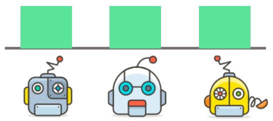

Speaker’s Utterance: u_t
Self-Conscious Speaker: S_1^t
$\propto L_0^t(i|h,u_{\le t},p_t)^\alpha * S_0^t(u_t|i,h,u_{<t})$
Imaginary Listener: $L_0^t(i|u_{\le t},h,p_t)$
Base Speaker: $S_0^t(u_t|i,h,u_{<t})$
$p_{t+1}(i)$
Learned Distractor Personas: i'
Dialogue History: h
Persona: i

Components of the Self-Conscious Speaker S_1

A Recursive Process in Bayesian Fashion

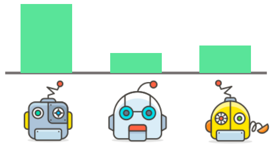
- *A base speaker* *(no self consciousness)*

$$S_0^t(u_t \mid i, h, u_{<t})$$



- *An imaginary listener*

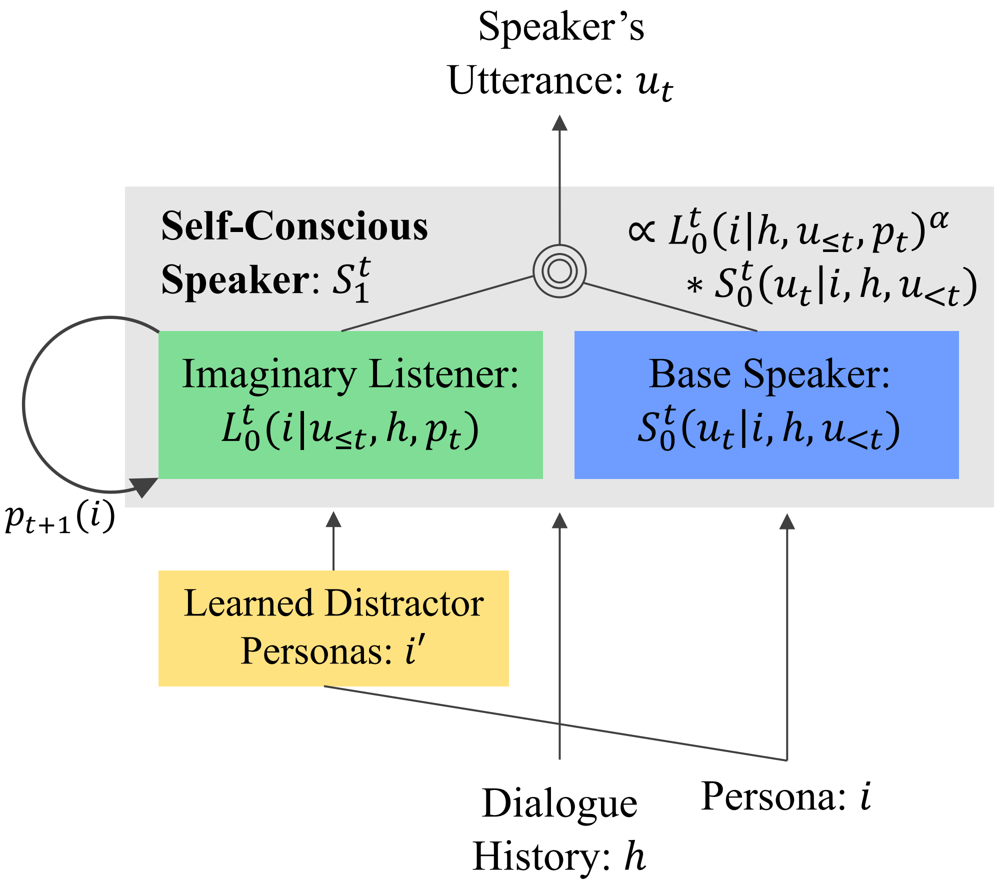
$$L_0^t(i \mid h, u_{\leq t}, p_t) \propto \frac{S_0^t(u_t \mid i, h, u_{<t})^\beta \cdot p_t(i)}{\sum_{i' \in \mathrm{I}} S_0^t(u_t \mid i, h, u_{<t})^\beta \cdot p_t(i')}$$



- *The **self conscious** speaker*

$$S_1^t(u_t \mid i, h, u_{<t})$$

$$\propto L_0^t(i \mid h, u_{\leq t}, p_t)^\alpha \cdot S_0^t(u_t \mid i, h, u_{<t})$$



Speaker's Utterance: u_t
Self-Conscious Speaker: S_1^t
$\propto L_0^t(i|h, u_{\leq t}, p_t)^\alpha$
$* S_0^t(u_t|i, h, u_{<t})$
Imaginary Listener: $L_0^t(i|u_{\leq t}, h, p_t)$
Base Speaker: $S_0^t(u_t|i, h, u_{<t})$
$p_{t+1}(i)$
Learned Distractor Personas: i'
Dialogue History: h
Persona: i

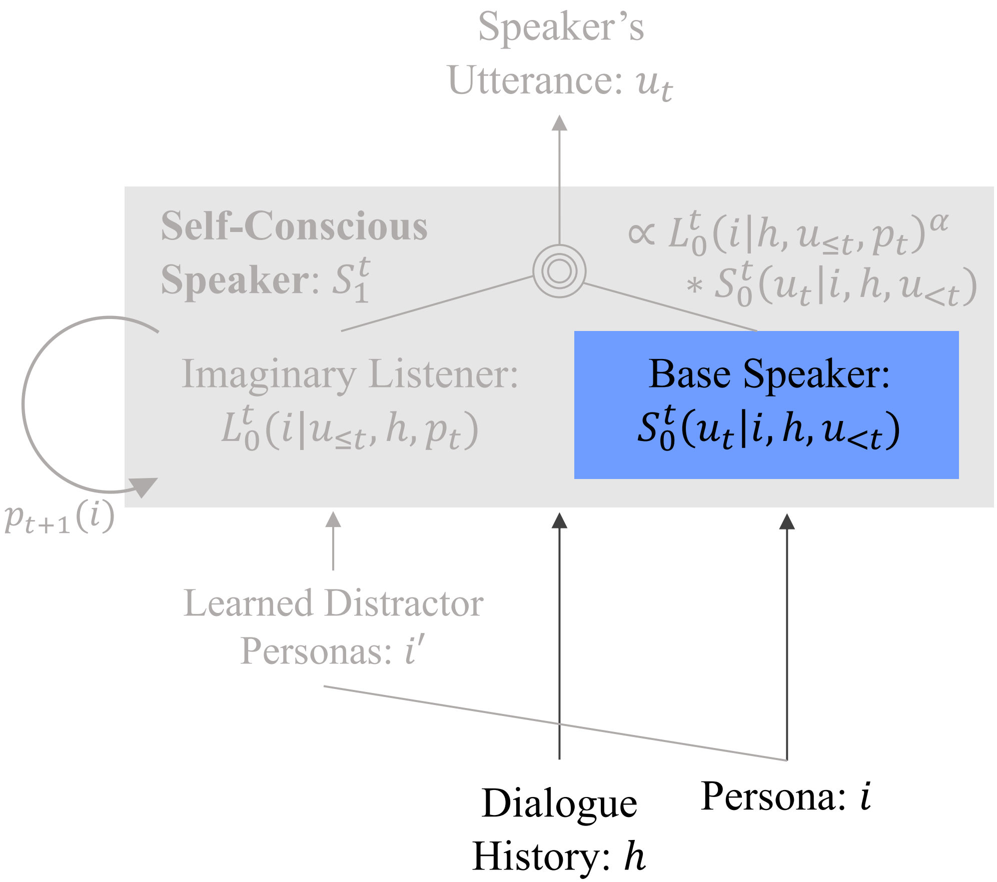
Base Speaker S_0

Any pretrained generative dialogue model
= Prior distribution

- *A base speaker* *(no self consciousness)*

$S_0^t(u_t \mid i, h, u_{<t})$

⇨ Generating one token at a time



Speaker's Utterance: u_t
Self-Conscious Speaker: S_1^t
$\propto L_0^t(i|h, u_{\leq t}, p_t)^\alpha * S_0^t(u_t|i, h, u_{<t})$
Imaginary Listener: $L_0^t(i|u_{\leq t}, h, p_t)$
Base Speaker: $S_0^t(u_t|i, h, u_{<t})$
$p_{t+1}(i)$
Learned Distractor Personas: i'
Dialogue History: h
Persona: i

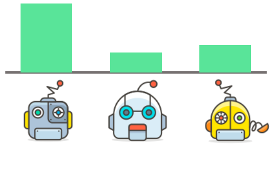
Imaginary Listener L_0

The likelihood of the given persona

Accumulative World Prior

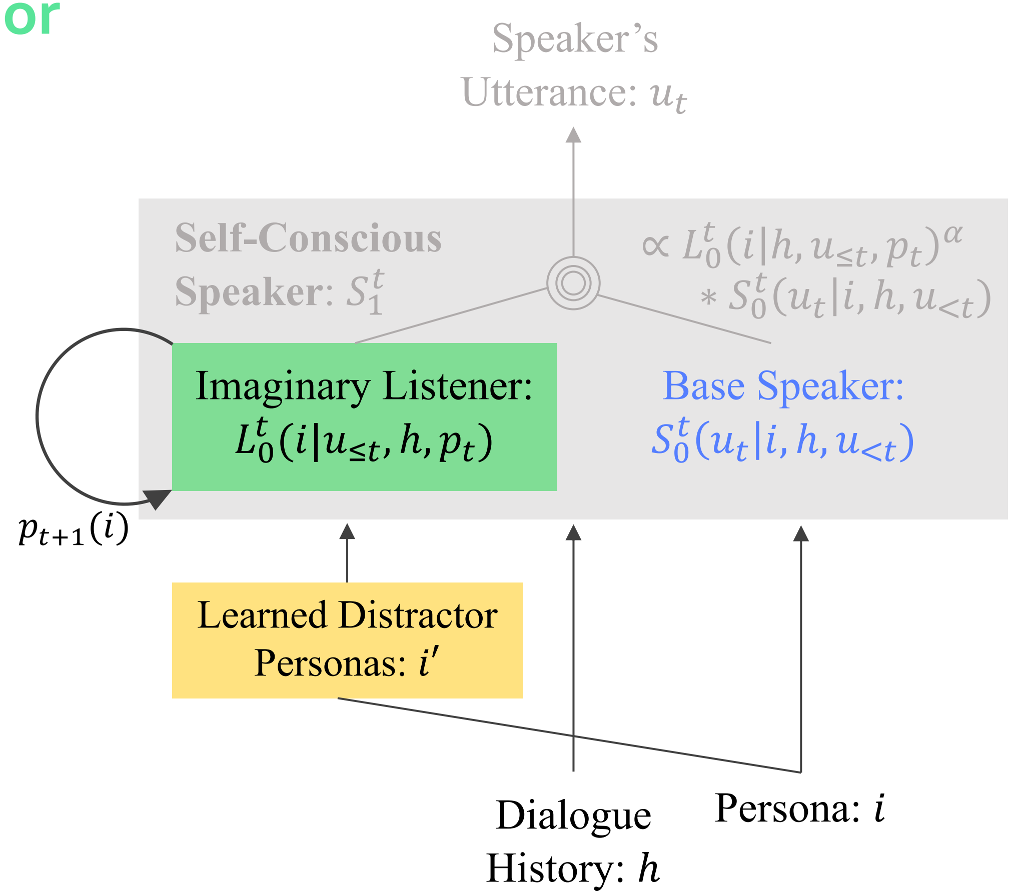
- *An imaginary listener*

$$L_0^t(i\,|h, u_{\le t}, p_t) \propto \frac{S_0^t(u_t\,|i, h, u_{<t})^\beta \cdot p_t(i)}{\sum_{i'\in \mathrm{I}} S_0^t(u_t\,|i, h, u_{<t})^\beta \cdot p_t\,(i')}$$



World I: given persona + distractors

Learned with Life-long Memory Networks



Speaker's Utterance: u_t

Self-Conscious Speaker: S_1^t

$\propto L_0^t(i|h, u_{\le t}, p_t)^\alpha$
$* \; S_0^t(u_t|i, h, u_{<t})$

Imaginary Listener: $L_0^t(i|u_{\le t}, h, p_t)$

Base Speaker: $S_0^t(u_t|i, h, u_{<t})$

$p_{t+1}(i)$

Learned Distractor Personas: i'

Dialogue History: h

Persona: i

- Note:

Use L_0 and β value less than 1 to prevent losing the cumulative information.
Previous work using L_1 reported indifference with using a uniform prior.

Kaiser et al. 2017. Learning to Remember Rare Events. *ICLR*

Cohn-Gordon et al. 2018. Pragmatically Informative Image Captioning With Character-Level Inference. *NAACL-HLT*

Self-Conscious Speaker S_1

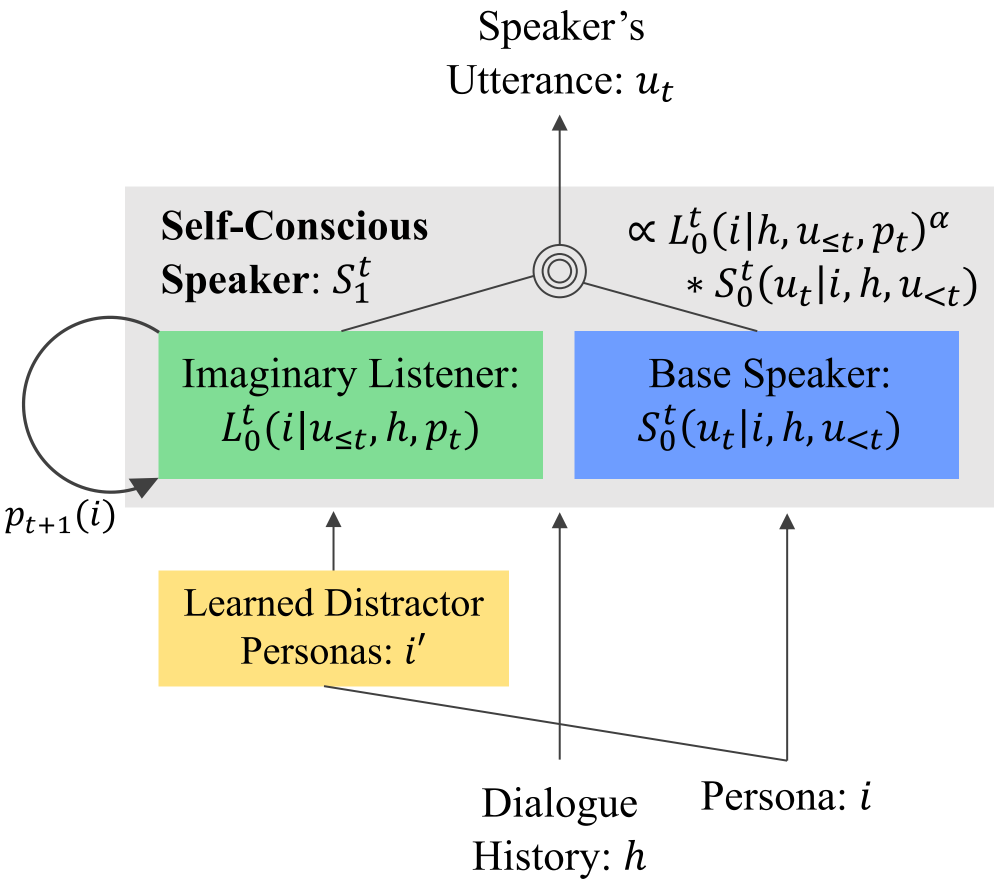
The posterior distribution

- *The **self conscious** speaker*

$$S_1^t(u_t \mid i, h, u_{<t})$$

$$\propto L_0^t(i \mid h, u_{\leq t}, p_t)^{\alpha} \cdot S_0^t(u_t \mid i, h, u_{<t})$$

Intensity of Self-consciousness
= Controlling the amount of the listener's information


Speaker's Utterance: u_t
Self-Conscious Speaker: S_1^t
$\propto L_0^t(i|h, u_{\leq t}, p_t)^{\alpha}$
$* S_0^t(u_t|i, h, u_{<t})$
Imaginary Listener: $L_0^t(i|u_{\leq t}, h, p_t)$
Base Speaker: $S_0^t(u_t|i, h, u_{<t})$
$p_{t+1}(i)$
Learned Distractor Personas: i'
Dialogue History: h
Persona: i

Experiments:
Dialogue NLI Evaluation Set
PersonaChat
Human Evaluation

Welleck et al. 2019. Dialogue Natural Language Inference. *ACL*

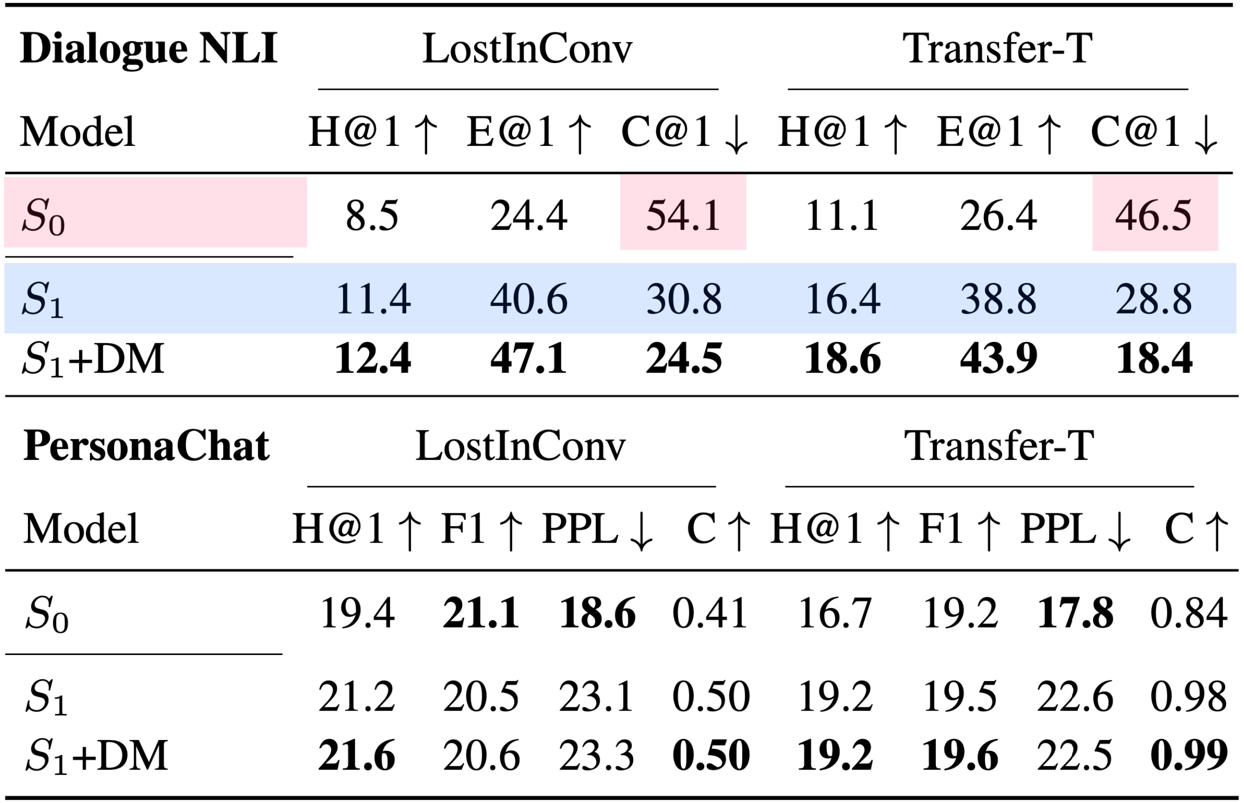
Zhang et al. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *ACL*

Results on Dialogue NLI

S_0: Base speaker model: **Lost In Conversation** & **Transfer Transfo**
S_1: Self-conscious speaker
+DM: Distractor Memory

Task:
31 candidate utterances given.
(1 ground-truth, 10 entailing, 10 neutral, 10 contradicting utterance)
The model selects the best utterance by perplexity

The proportion of selecting
Ground-truth (**H**its**@1**)
Entailing utterance (**E**ntail**@1**)
Contradicting utterance (**C**ontradict**@1**)



Dialogue NLI	LostInConv			Transfer-T		
Model	H@1 ↑	E@1 ↑	C@1 ↓	H@1 ↑	E@1 ↑	C@1 ↓
S_0	8.5	24.4	54.1	11.1	26.4	46.5
S_1	11.4	40.6	30.8	16.4	38.8	28.8
S_1+DM	**12.4**	**47.1**	**24.5**	**18.6**	**43.9**	**18.4**

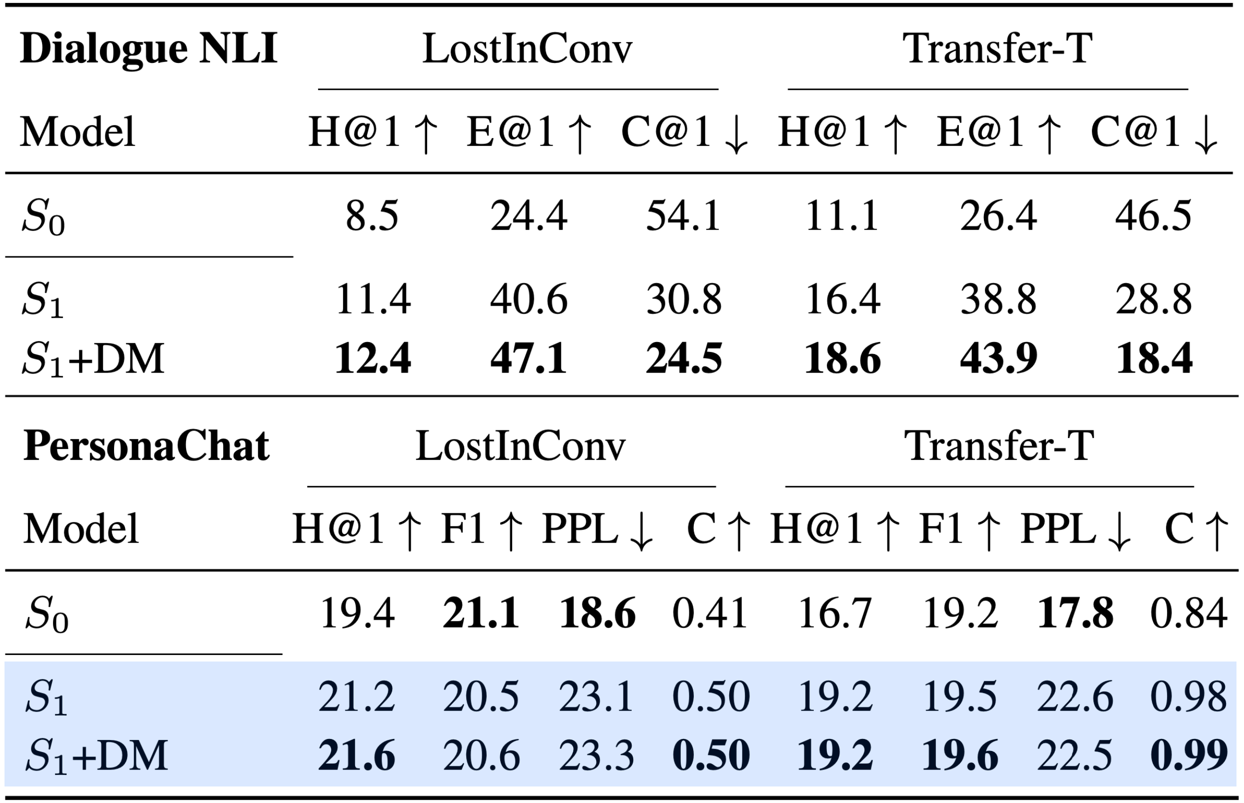
PersonaChat	LostInConv				Transfer-T			
Model	H@1 ↑	F1 ↑	PPL ↓	C ↑	H@1 ↑	F1 ↑	PPL ↓	C ↑
S_0	19.4	**21.1**	**18.6**	0.41	16.7	19.2	**17.8**	0.84
S_1	21.2	20.5	23.1	0.50	19.2	19.5	22.6	0.98
S_1+DM	**21.6**	20.6	23.3	**0.50**	**19.2**	**19.6**	22.5	**0.99**

Alexander Tselousov and Sergey Golovanov. 2019. Lost In Conversation.
Wolf et al. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv*

Results on PersonaChat

S_0: Base speaker model: **Lost In Conversation** & **Transfer Transfo**
S_1: Self-conscious speaker
+DM: Distractor Memory

C: consistency score,
evaluation with pretrained NLI model



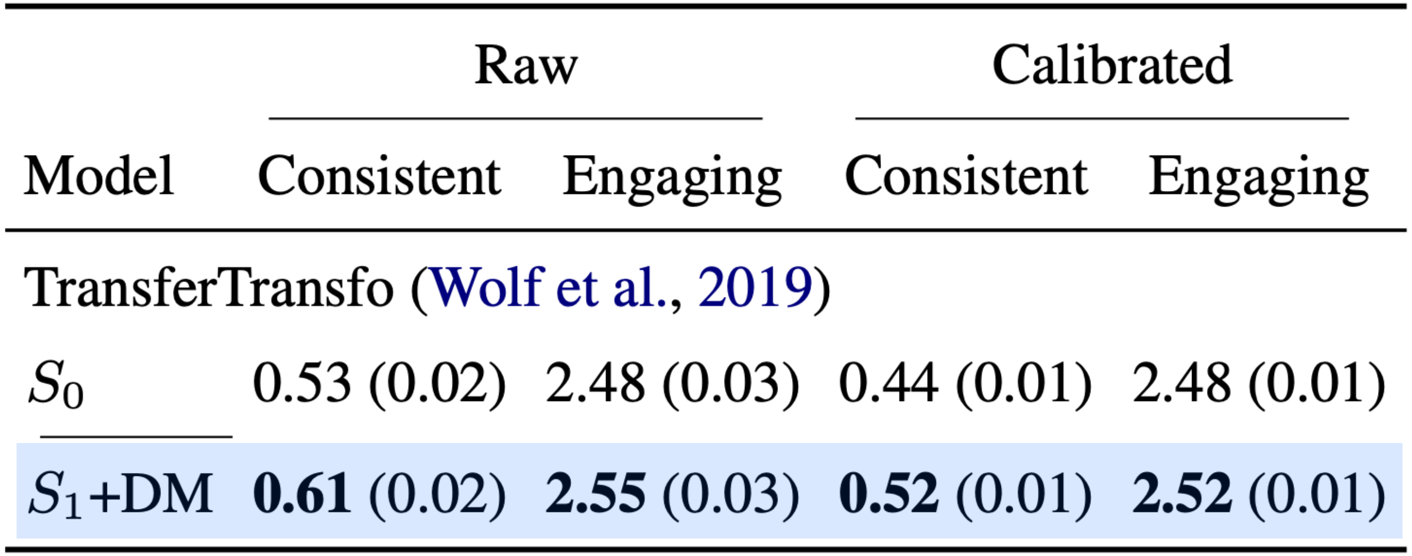
Dialogue NLI	LostInConv			Transfer-T		
Model	H@1 ↑	E@1 ↑	C@1 ↓	H@1 ↑	E@1 ↑	C@1 ↓
S_0	8.5	24.4	54.1	11.1	26.4	46.5
S_1	11.4	40.6	30.8	16.4	38.8	28.8
S_1+DM	**12.4**	**47.1**	**24.5**	**18.6**	**43.9**	**18.4**

PersonaChat	LostInConv				Transfer-T			
Model	H@1 ↑	F1 ↑	PPL ↓	C ↑	H@1 ↑	F1 ↑	PPL ↓	C ↑
S_0	19.4	**21.1**	**18.6**	0.41	16.7	19.2	**17.8**	0.84
S_1	21.2	20.5	23.1	0.50	19.2	19.5	22.6	0.98
S_1+DM	**21.6**	20.6	23.3	**0.50**	**19.2**	**19.6**	22.5	**0.99**

Madotto et al. 2019. Personalizing Dialogue Agents via Meta-Learning. *ACL*

Results on Human Evaluation

Consistency: *Is the response consistent?*
Engagingness: *How much do you like the response?*
on TransferTransfo model



	Raw		Calibrated	
Model	Consistent	Engaging	Consistent	Engaging
TransferTransfo (Wolf et al., 2019)				
S_0	0.53 (0.02)	2.48 (0.03)	0.44 (0.01)	2.48 (0.01)
S_1+DM	**0.61** (0.02)	**2.55** (0.03)	**0.52** (0.01)	**2.52** (0.01)

Numbers in parentheses are standard error
We also report Bayesian calibrated scores to remove evaluator bias

Kulikov et al. 2019. Importance of Search and Evaluation Strategies in Neural Dialogue Modeling. *INLG*

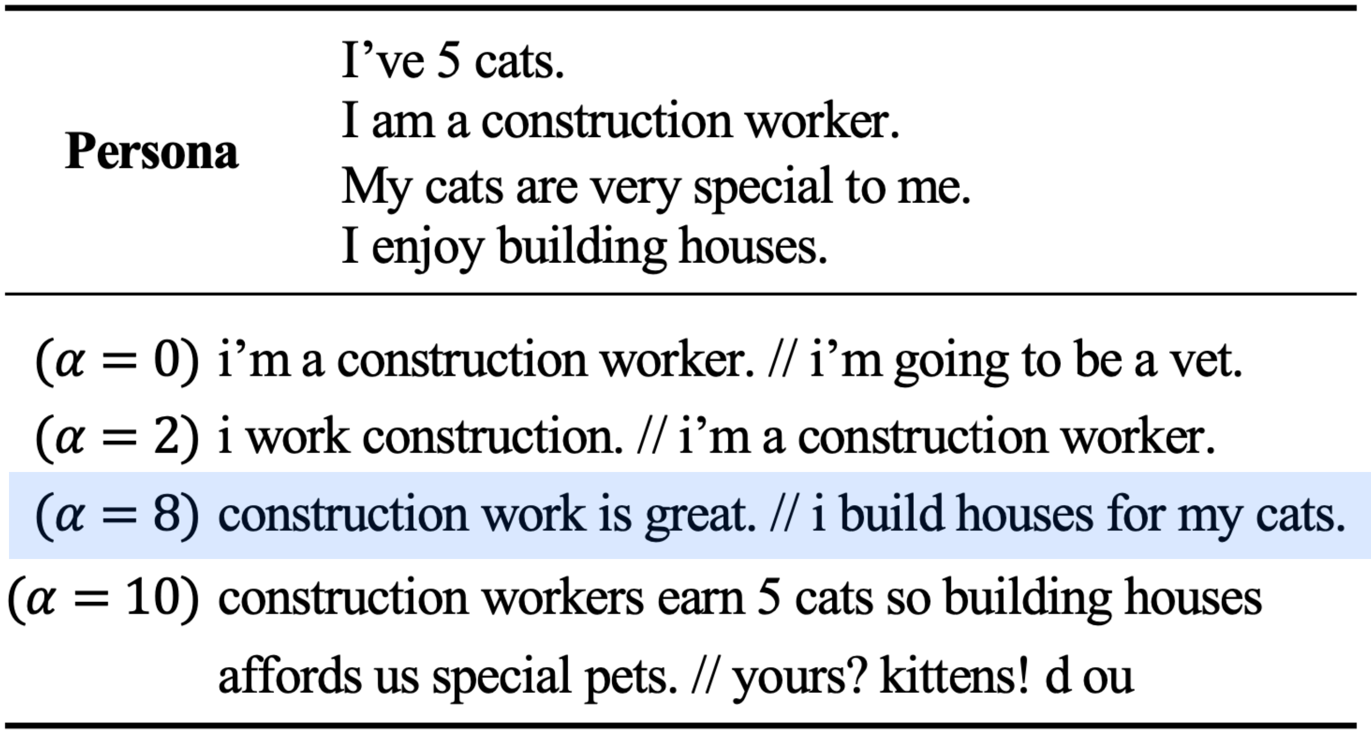
Controlling the Self-conscious agent: α and β

α controls the degree of copying the given condition text (=persona)

Appropriate value allows the condition text to blend smoothly in the generation

- *The **self conscious** speaker*

$$S_1^t(u_t \mid i, h, u_{<t}) \propto L_0^t(i \mid h, u_{\le t}, p_t)^{\alpha} \cdot S_0^t(u_t \mid i, h, u_{<t})$$



Persona	I've 5 cats. I am a construction worker. My cats are very special to me. I enjoy building houses.
$(\alpha = 0)$	i'm a construction worker. // i'm going to be a vet.
$(\alpha = 2)$	i work construction. // i'm a construction worker.
$(\alpha = 8)$	construction work is great. // i build houses for my cats.
$(\alpha = 10)$	construction workers earn 5 cats so building houses affords us special pets. // yours? kittens! d ou

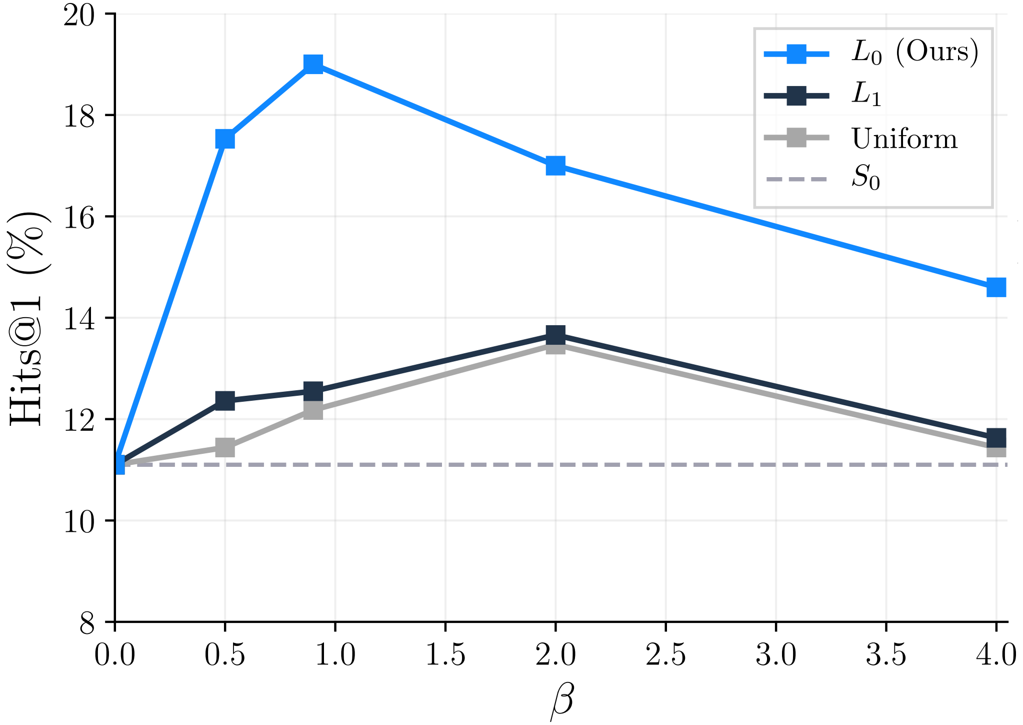
β and **World prior** $p_t(i)$

Value equal to 1 or slightly less updating the world prior with L_0 is appropriate for incremental decoding

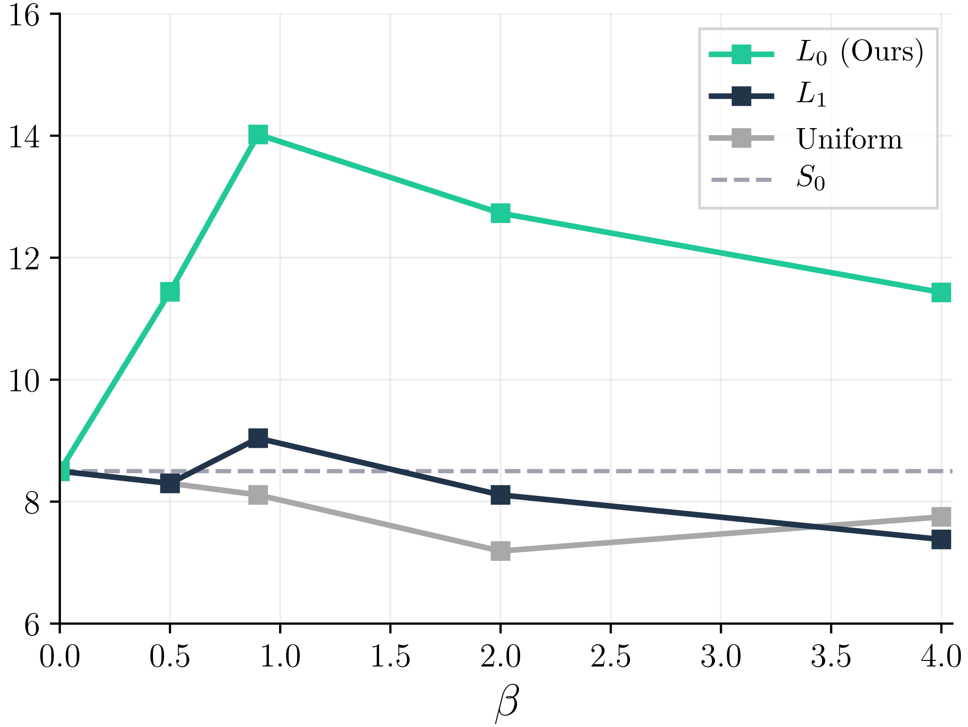
- *An imaginary listener*

$$L_0^t(i \mid h, u_{\le t}, p_t) \propto \frac{S_0^t(u_t \mid i, h, u_{<t})^{\beta} \cdot p_t(i)}{\sum_{i' \in \mathrm{I}} S_0^t(u_t \mid i, h, u_{<t})^{\beta} \cdot p_t(i')}$$

TransferTransfo



LostInConv



Concluding Remarks

- Introduced an ***unsupervised*** method for improving consistency inspired by social cognition and pragmatics

  → **Requiring no additional annotations nor external models**

- Further extended the Rational Speech Acts framework

  → **Learning to provide distractors and different update for world prior**

- Extensive experiments on Dialogue NLI, PersonaChat and Human Evalution

  → **Significantly reduced contradiction and improved ground-truth accuracy**

Thank you

Images and Icons are from Facebook ConvAI2 challenge, Nhor Phai, and Vincent Le Moign.