



EMNLP
2022



BotsTalk: Machine-sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets



Minju Kim



Chaehyeong Kim



Yongho Song



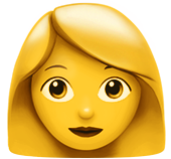
Seung-won Hwang



Jinyoung Yeo

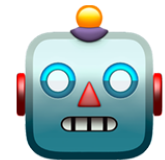
Open-domain Chatbots with Desirable Qualities for Conversations

Open-domain Chatbots with *Personality*



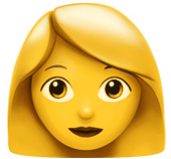
My uncle gave me a job. I was considering college, but I am good now. You?

I am almost done. I only have two years left in law school.



Open-domain Chatbots with *Personality*

PersonaChat (2016)

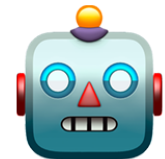


My uncle gave me a job. I was considering college, but I am good now. You?

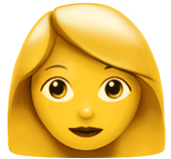
Persona

I have got two more years in college; I study law

I am almost done. I only have two years left in law school.

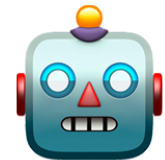


Open-domain Chatbots with *Knowledge*



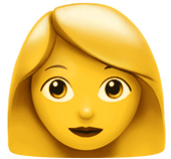
I love ice cream as much as anyone. I especially like Gelato, foreign ice cream!

Have you heard of bacon ice cream? They add bacon and egg custard to the freezing mixture!



Open-domain Chatbots with *Knowledge*

Wizard of Wikipedia (2019)



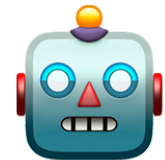
I love ice cream as much as anyone. I especially like Gelato, foreign ice cream!

Topic &
Article

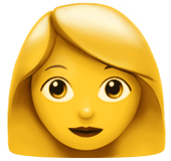


Ice cream; Bacon ice cream is an ice cream generally created by adding bacon to egg custard and freezing the mixture.

Have you heard of bacon ice cream? They add bacon and egg custard to the freezing mixture!

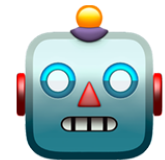


Open-domain Chatbots with *Empathy*



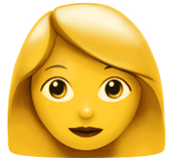
I've been hearing some strange noises around the house at night.

Oh no! That's scary!
What do you think it is?



Open-domain Chatbots with *Empathy*

Empathetic Dialogues (2019)

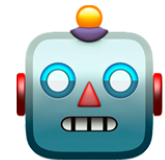


I've been hearing some strange noises around the house at night.

Situation
& Emotion

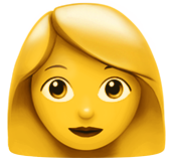
I've been hearing noise around the house at night; Afraid

Oh no! That's scary!
What do you think it is?



What's next?

Human conversation requires a variety of conversational skills!

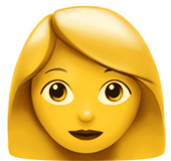


Do you know much about aeros?
They have excellent fit and beauty.

Personal utterance

I have a hard time finding
it but they are great shoes.

Knowledgeable utterance

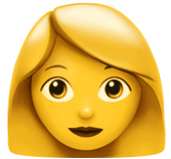


I hope you can find ones
that are comfortable to you.

Empathetic utterance

What's next?

Human conversation requires a variety of conversational skills!

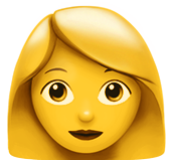


Do you know much about aeros?
They have excellent fit and beauty.

Personal utterance

I have a hard time finding
it but they are great shoes.

Knowledgeable utterance



I hope you can find ones
that are comfortable to you.

Empathetic utterance



Towards multi-skill
dialogue systems

Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.	Blended Skill Talk (BST)
Wizard of Wikipedia topic: Video game design Previous utterances (shown to speakers): U: What video games do you like to play? G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget		
Actual utterances: U: Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) (G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.") G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K) U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K) G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S) U: I use games to relax after a stressful day, the small escape is relaxing. (PB) (G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!") G: I enjoy a good gaming session after a hard day at work as well. (PB) U: What other hobbies does your son have? (PB) G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB) U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB) G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. (PB) U: Sounds nice. Are they new or vintage? (E) G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)		

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.
Wizard of Wikipedia topic: Video game design Previous utterances (shown to speakers): U: What video games do you like to play? G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget	
Actual utterances: U Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) <i>(G selected the wow suggestion. Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.)</i> G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K) U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K) G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S) U: I use games to relax after a stressful day, the small escape is relaxing. (PB) <i>(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")</i> G: I enjoy a good gaming session after a hard day at work as well. (PB) U: What other hobbies does your son have? (PB) G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB) U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB) G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. (PB) U: Sounds nice. Are they new or vintage? (E) G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)	

Blended Skill Talk (BST)

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.
Wizard of Wikipedia topic: Video game design Previous utterances (shown to speakers): U: What video games do you like to play? G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget	
Actual utterances: U Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) <i>(G selected the wow suggestion. Indeed, Some games games are purposely designed to be a work of a persons creative expression many though have been challenged as works of art by some critics")</i> G Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K) U: video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K) G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S) U: I use games to relax after a stressful day, the small escape is relaxing. (PB) <i>(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")</i> G: I enjoy a good gaming session after a hard day at work as well. (PB) U: What other hobbies does your son have? (PB) G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB) U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB) G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. (PB) U: Sounds nice. Are they new or vintage? (E) G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)	

Blended Skill Talk (BST)

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.
Wizard of Wikipedia topic: Video game design Previous utterances (shown to speakers): U: What video games do you like to play? G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget	
Actual utterances: U Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) <i>(G selected the wow suggestion. "Indeed, Some games games are purposely designed to be a work of a persons creative expression many though have been challenged as works of art by some critics")</i> G Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K) U video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K) G Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S) U: I use games to relax after a stressful day, the small escape is relaxing. (PB) <i>(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")</i> G: I enjoy a good gaming session after a hard day at work as well. (PB) U: What other hobbies does your son have? (PB) G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB) U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB) G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. (PB) U: Sounds nice. Are they new or vintage? (E) G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)	

Blended Skill Talk (BST)

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

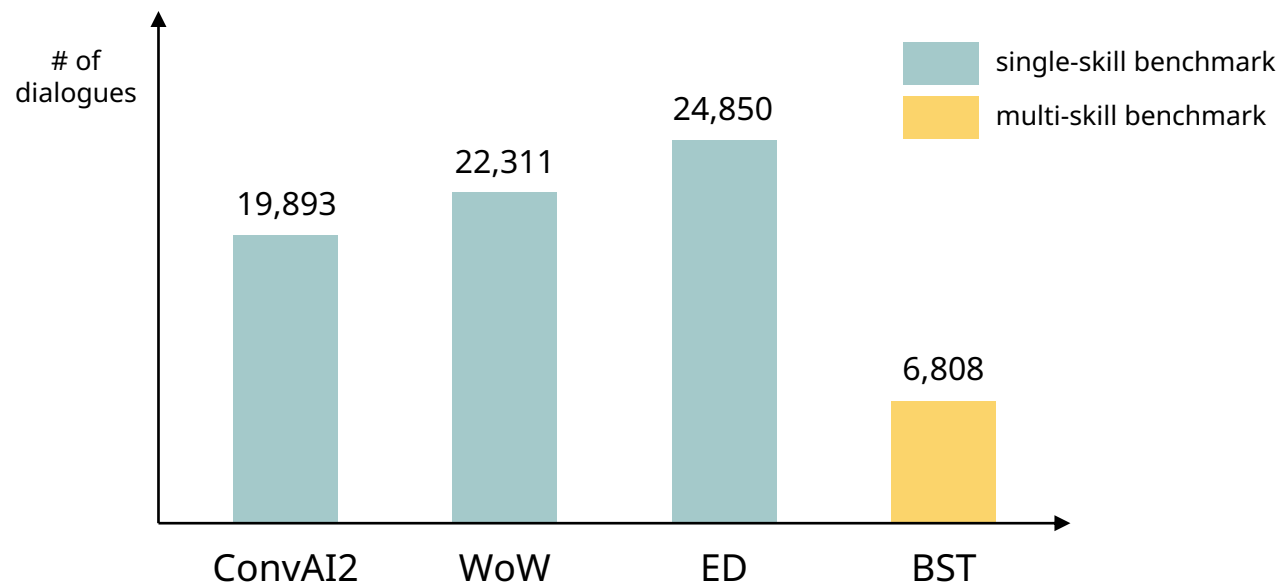
Persona for Unguided Speaker: My son plays on the local football team. I design video games for a living.	Persona for Guided Speaker: My eyes are green. I wear glasses that are cateye.
Wizard of Wikipedia topic: Video game design Previous utterances (shown to speakers): U: What video games do you like to play? G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget	
Actual utterances: U Exactly! I think many people fail to notice how beautiful the art of video games can be. (PB) <i>(G selected the wow suggestion. "Indeed, Some games games are purposely designed to be a work of a persons creative expression many though have been challenged as works of art by some critics")</i> G Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (K) U video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (K) G Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (S) U: I use games to relax after a stressful day, the small escape is relaxing. (PB) <i>(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")</i> G: I enjoy a good gaming session after a hard day at work as well. (PB) U: What other hobbies does your son have? (PB) G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (PB) U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (PB) G: I wish I could play football. But I wear this cateye glasses and they would break if I tried. (PB) U Sounds nice. Are they new or vintage? (E) G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (S)	

Blended Skill Talk (BST)

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

Limitations of crowdsourcing

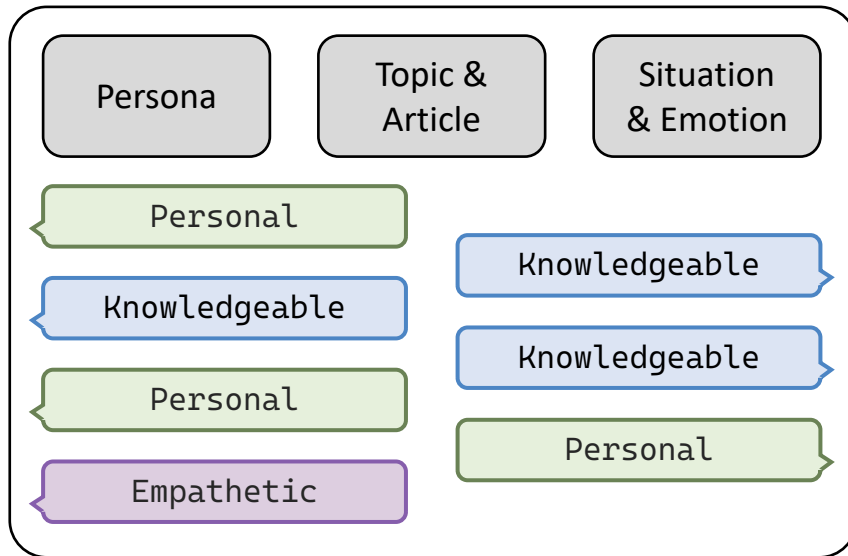
Limited scale of dataset



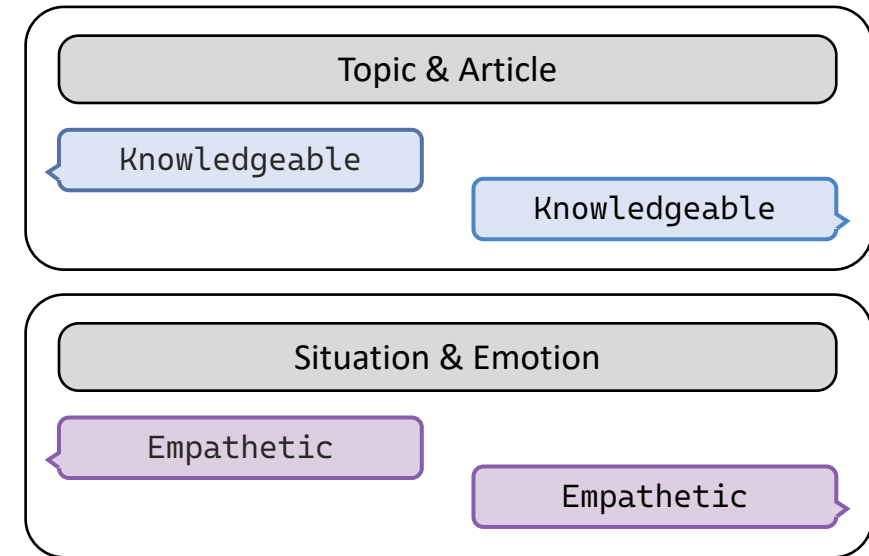
Limitations of crowdsourcing

Limited scale of dataset

➡ Multi-skill dialogues are much more *complex* than single-skill dialogues
Why?



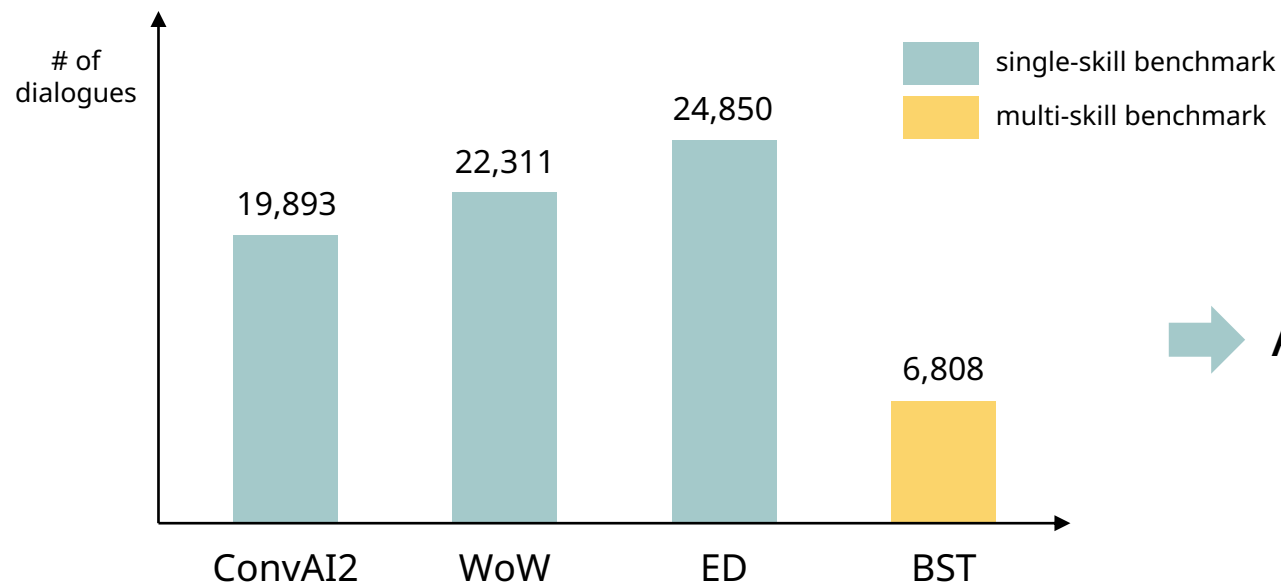
Multi-skill dialogue dataset



Single-skill dialogue datasets

Limitations of crowdsourcing

Limited scale of dataset



➡ Automatic data construction is required!

Limitations of crowdsourcing

What if more skill types appear?

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

**Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Orianna Demasi³, Sahand Sabour¹, Yu Li³,
Zhou Yu⁴, Yong Jiang², Minlie Huang^{1†}**

¹The CoAI group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-118@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Oriana Riva³, Aihuang Hu⁴, Yu Li³,

Fine-grained skill type

¹The CoAI group, Tsinghua University, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-118@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Orianna P. ~~...~~ Li³,

Fine-grained skill type

¹The CoAI group, Tsinghua University, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-l18@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Image-Chat: Engaging Grounded Conversations

Kurt Shuster, Samuel Humeau, Antoine Bordes, Jason Weston

Facebook AI Research

{kshuster, samuelhumeau, abordes, jase}@fb.com

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Orianna Perle³, Aihuang Li³,

Fine-grained skill type

¹The CoAI group, Tsinghua University, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-l18@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Image-Chat: Engaging Grounded Conversations

Multi-modal as skill type

Samuel Humeau, Abordes, Jason Weston

Facebook AI Research

{kshuster, samuelhumeau, abordes, jase}@fb.com

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Oriana Riva³, and Aihuang Li³,

Fine-grained skill type

¹The CoAI group, Tsinghua University, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-l18@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Image-Chat: Engaging Grounded Conversations

Multi-modal as skill type

Samuel Humeau, Jason Weston

Facebook AI Research

{kshuster, samuelhumeau, abordes, jase}@fb.com



Scalable Framework

Scalable Transfer Learning with Expert Models

Joan Puigcerver*
Google Research

Carlos Riquelme*
Google Research

Basil Mustafa
Google Research

Cedric Renggli†
ETH Zurich

André Susano Pinto
Google Research

Sylvain Gelly
Google Research

Daniel Keysers
Google Research

Neil Houlsby
Google Research

Limitations of crowdsourcing

What if more skill types appear?

Towards Emotional Support Dialog Systems

Siyang Liu^{1,2*}, Chujie Zheng^{1*}, Orianna Perle³, and Liang He⁴

Fine-grained skill type

¹The CoAI group, Tsinghua University, State Key Lab of Intelligent Technology and Systems,

¹Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

²Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,

²Tsinghua University, Shenzhen, China

³University of California, Davis ⁴Columbia University

siyang-l18@mails.tsinghua.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Image-Chat: Engaging Grounded Conversations

Multi-modal as skill type

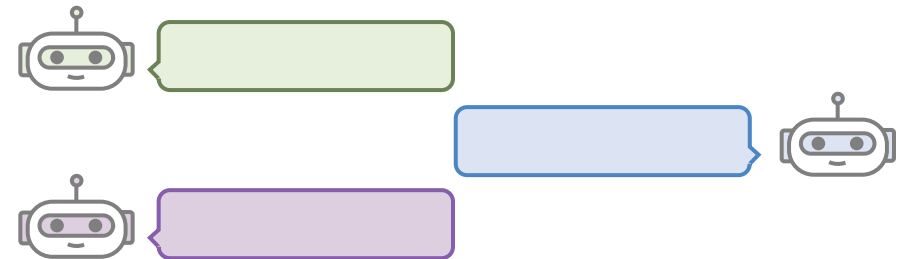
Samuel Humeau, Jason Weston

Facebook AI Research

{kshuster, samuelhumeau, abordes, jase}@fb.com



Scalable Framework

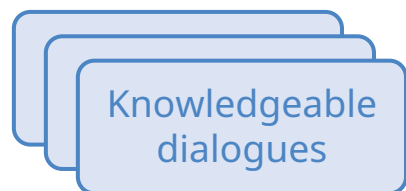


“ BotsTalk Framework ”

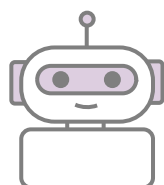
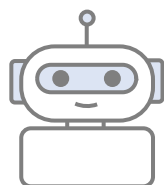
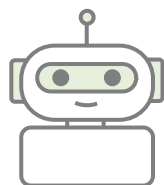
BotsTalk Framework

Goal

Single-skill
dialogue datasets

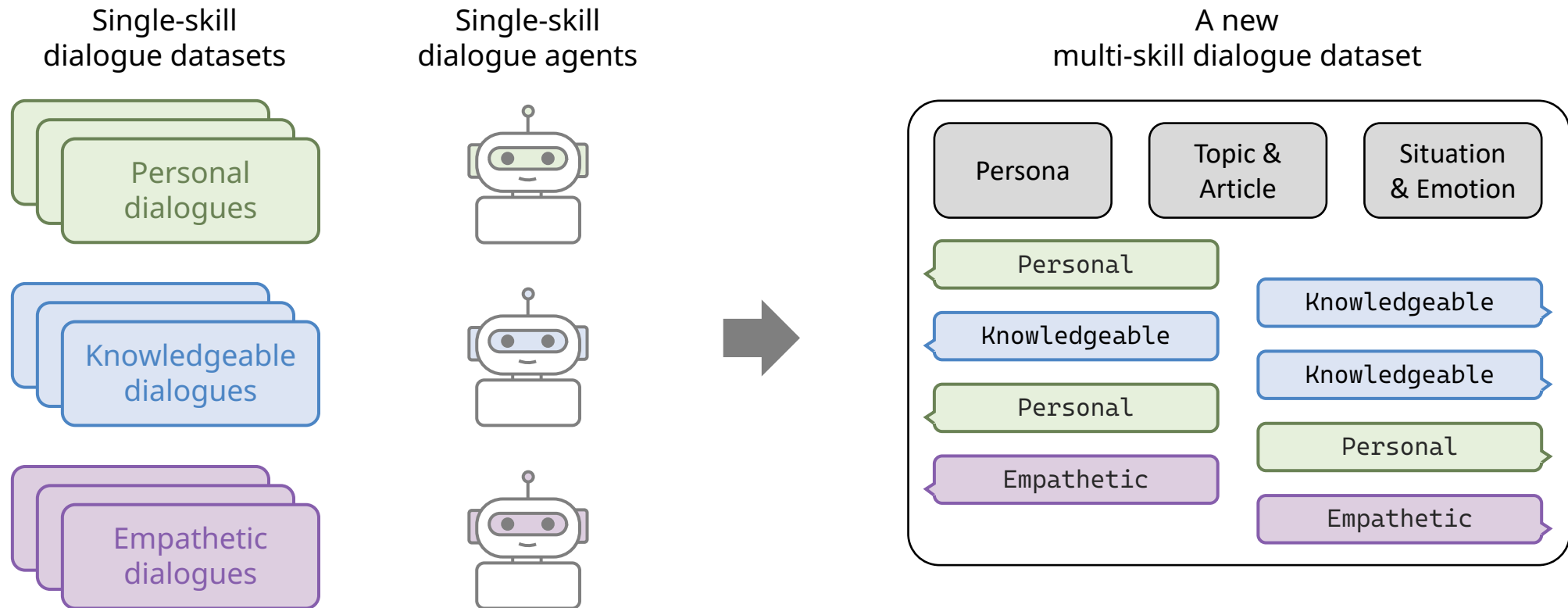


Single-skill
dialogue agents



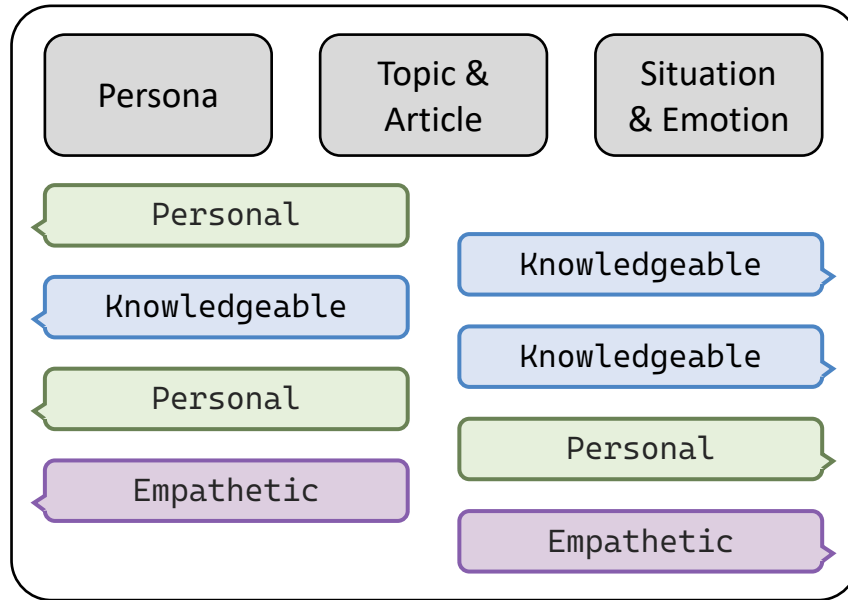
BotsTalk Framework

Goal



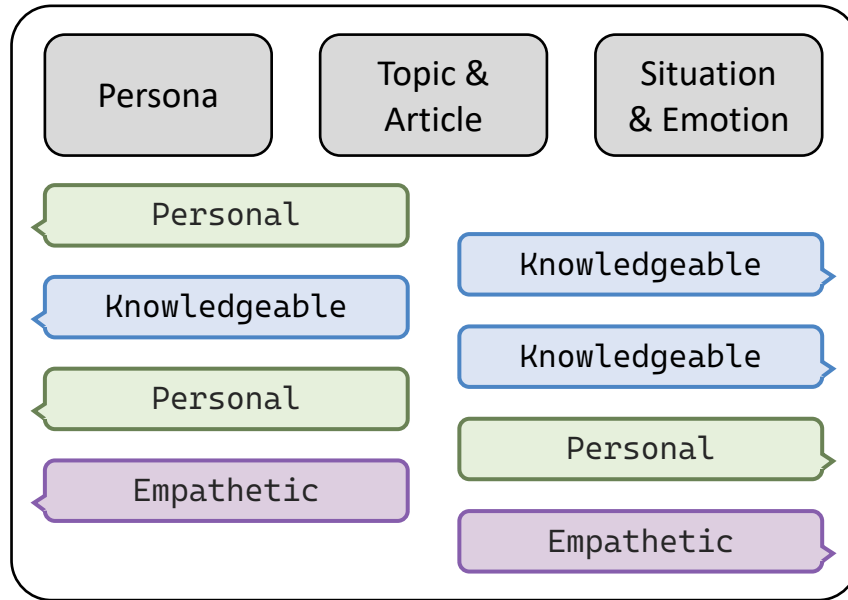
BotsTalk Framework

Desirable characteristics of multi-skill dialogue datasets



BotsTalk Framework

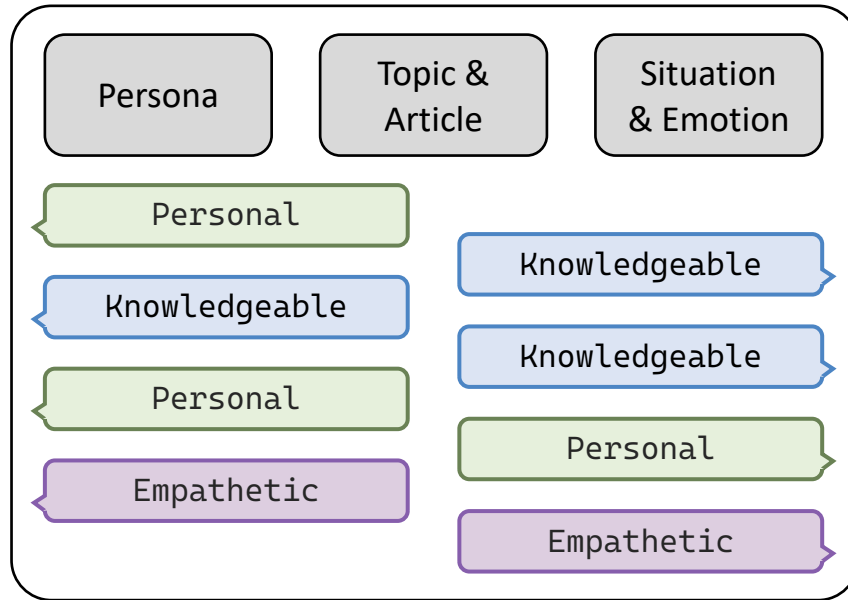
Desirable characteristics of multi-skill dialogue datasets



Requirements →

BotsTalk Framework

Desirable characteristics of multi-skill dialogue datasets



Requirements

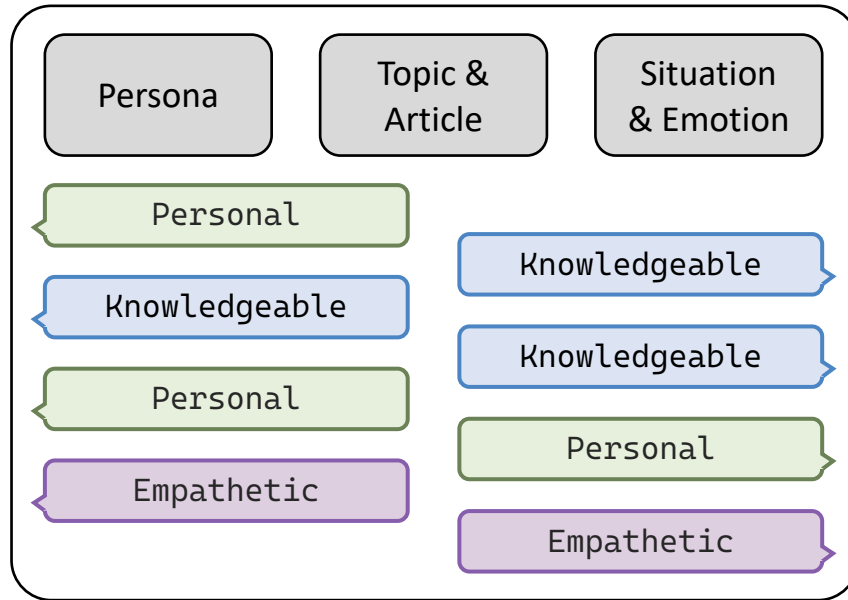


Skill blending

“ A dialogue model should learn to exhibit different dialogue skills in a conversation (Smith et al., 2020). ”

BotsTalk Framework

Desirable characteristics of multi-skill dialogue datasets



Requirements



Skill blending

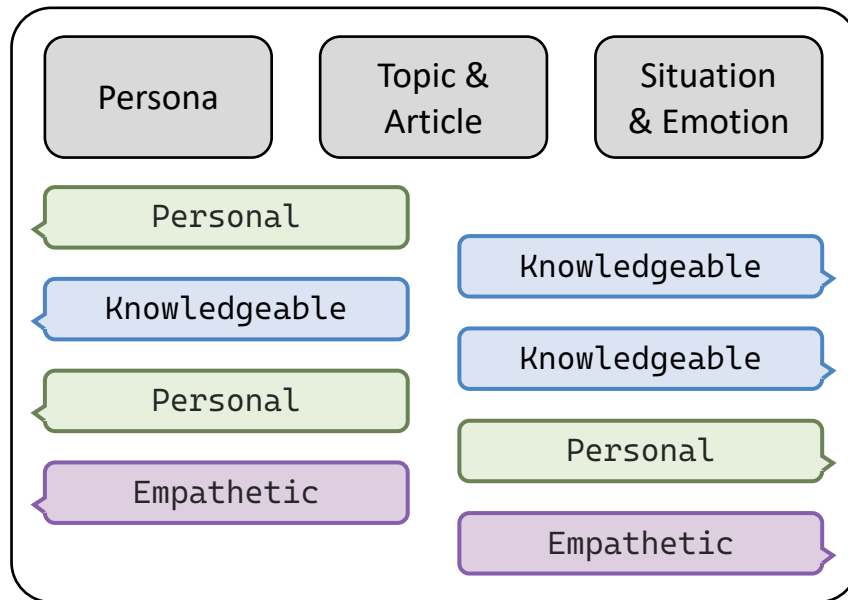
“ A dialogue model should learn to exhibit different dialogue skills in a conversation (Smith et al., 2020). ”

Skill grounding

“ A dialogue model should learn to maintain each dialogue skill when appropriate (Shazeer et al., 2017). ”

BotsTalk Framework

Desirable characteristics of multi-skill dialogue datasets



Skill blending

“ A dialogue model should learn to exhibit different dialogue skills in a conversation (Smith et al., 2020). ”

Skill grounding

“ A dialogue model should learn to maintain each dialogue skill when appropriate (Shazeer et al., 2017). ”

Generally, skill blending and skill grounding have a trade-off relationship (Madotto et al., 2021) as it is insufficient to represent both skill blending and grounding in a conversation of finite length.

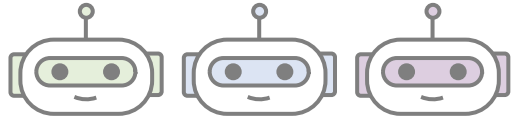
BotsTalk Framework

Participants in BotsTalk

BotsTalk Framework

Participants in BotsTalk

Skill agents

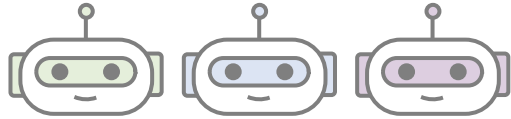


Multiple single-skill agents participate in a conversation to annotate the appropriate skill-grounded utterances to the dialogue.

BotsTalk Framework

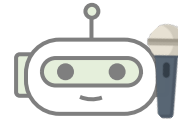
Participants in BotsTalk

Skill agents



Multiple single-skill agents participate in a conversation to annotate the appropriate skill-grounded utterances to the dialogue.

Active agent

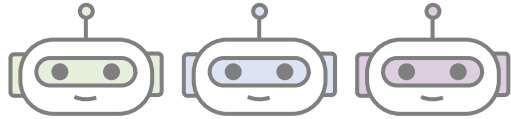


Only one skill agent is referred to as the "active agent", who currently has priority (mic 🎤) over the conversational flow. The active agent is willing to pass the mic to other agents if necessary.

BotsTalk Framework

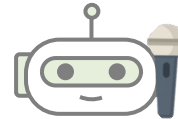
Participants in BotsTalk

Skill agents



Multiple single-skill agents participate in a conversation to annotate the appropriate skill-grounded utterances to the dialogue.

Active agent



Only one skill agent is referred to as the "active agent," who currently has priority (mic 🎤) over the conversational flow. The active agent is willing to pass the mic to other agents if necessary.

Moderator agent



Moderator agent is an omniscient oracle which controls the conversational flow and mediates the skill agents.

BotsTalk Framework

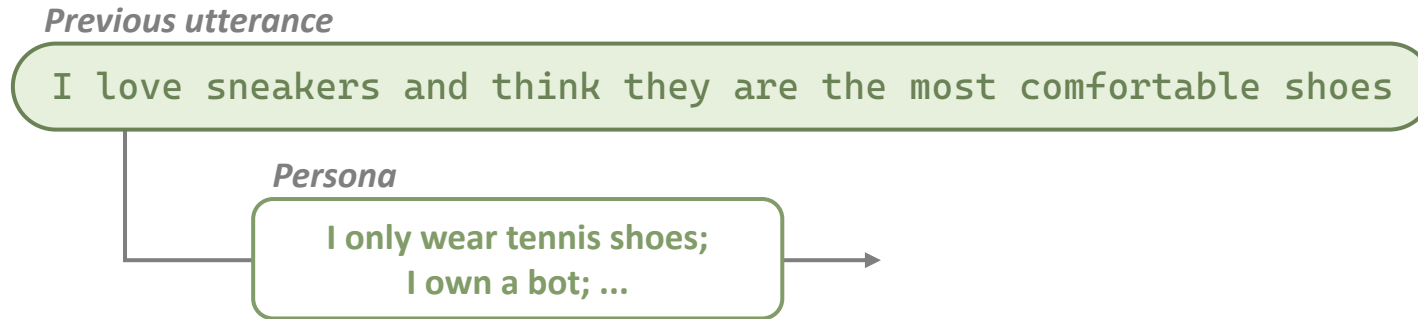
Phase 1: Simulate what to speak

Previous utterance

I love sneakers and think they are the most comfortable shoes

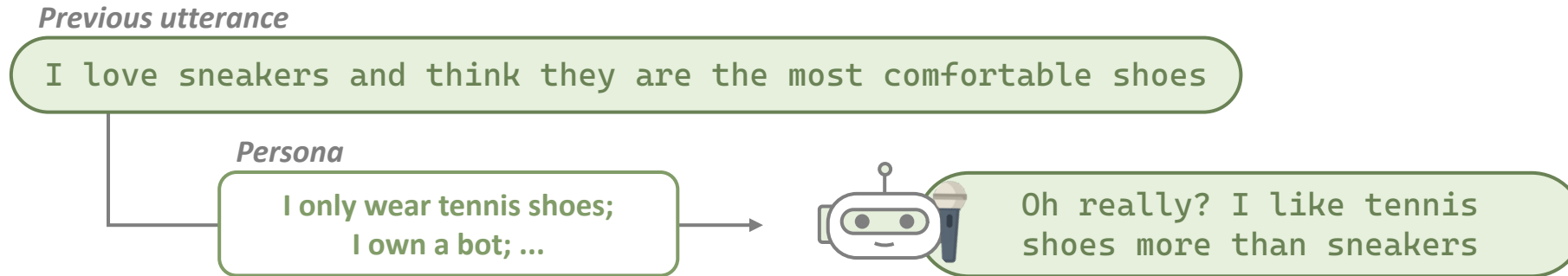
BotsTalk Framework

Phase 1: Simulate what to speak



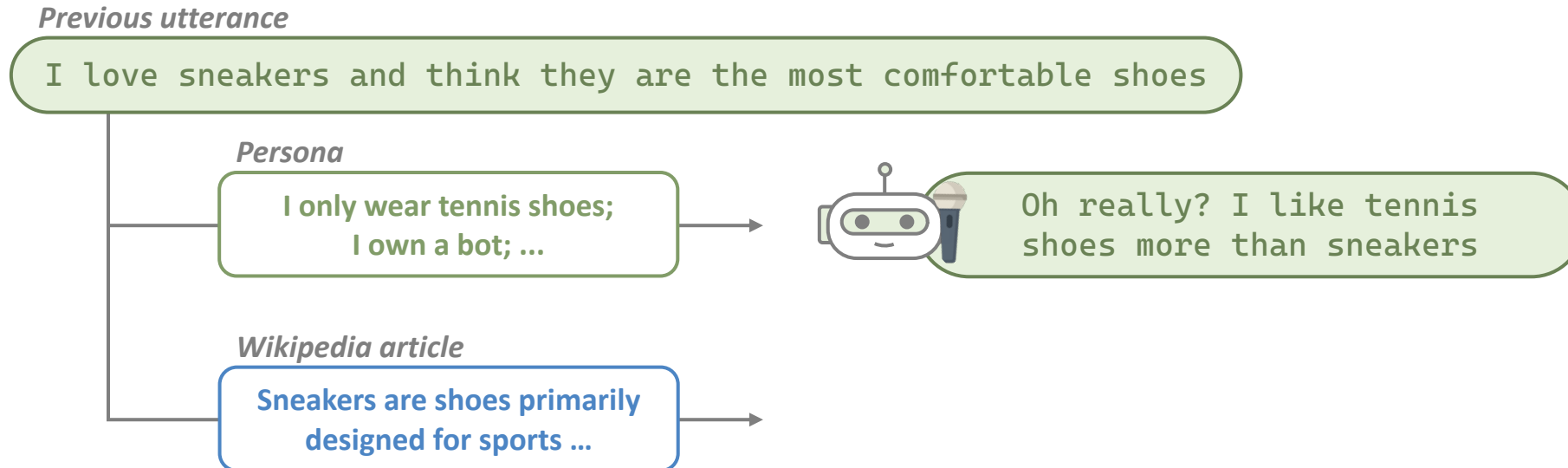
BotsTalk Framework

Phase 1: Simulate what to speak



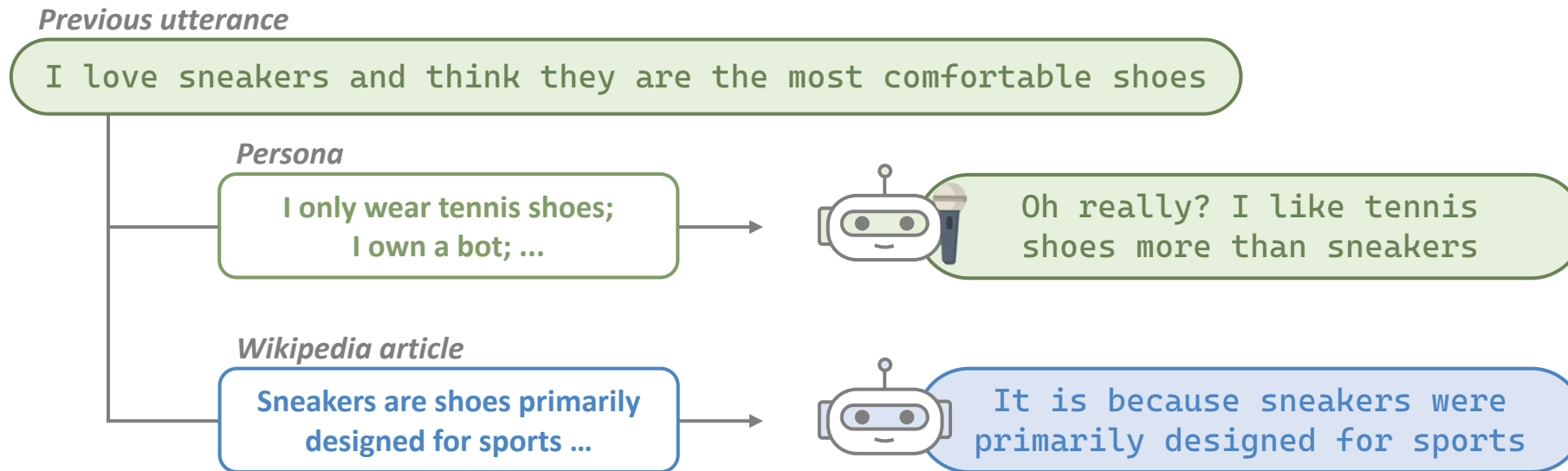
BotsTalk Framework

Phase 1: Simulate what to speak



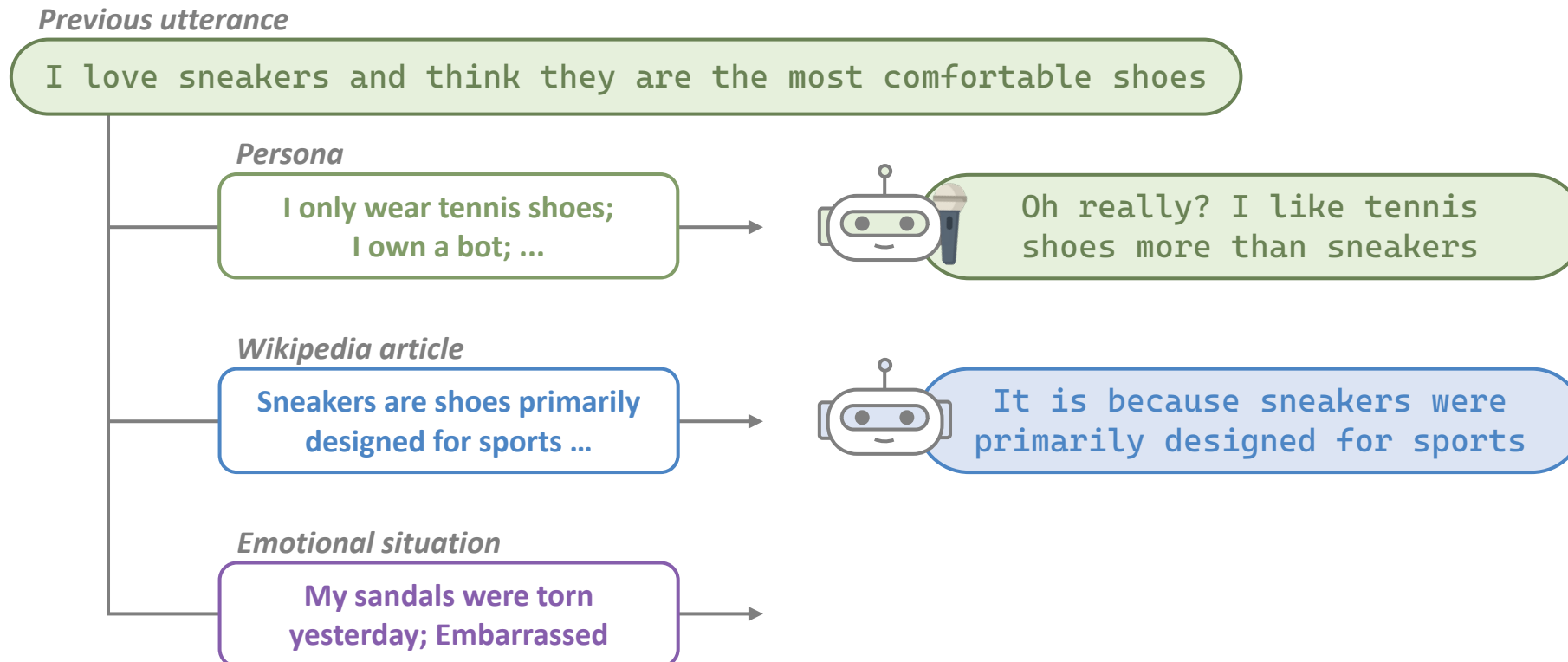
BotsTalk Framework

Phase 1: Simulate what to speak



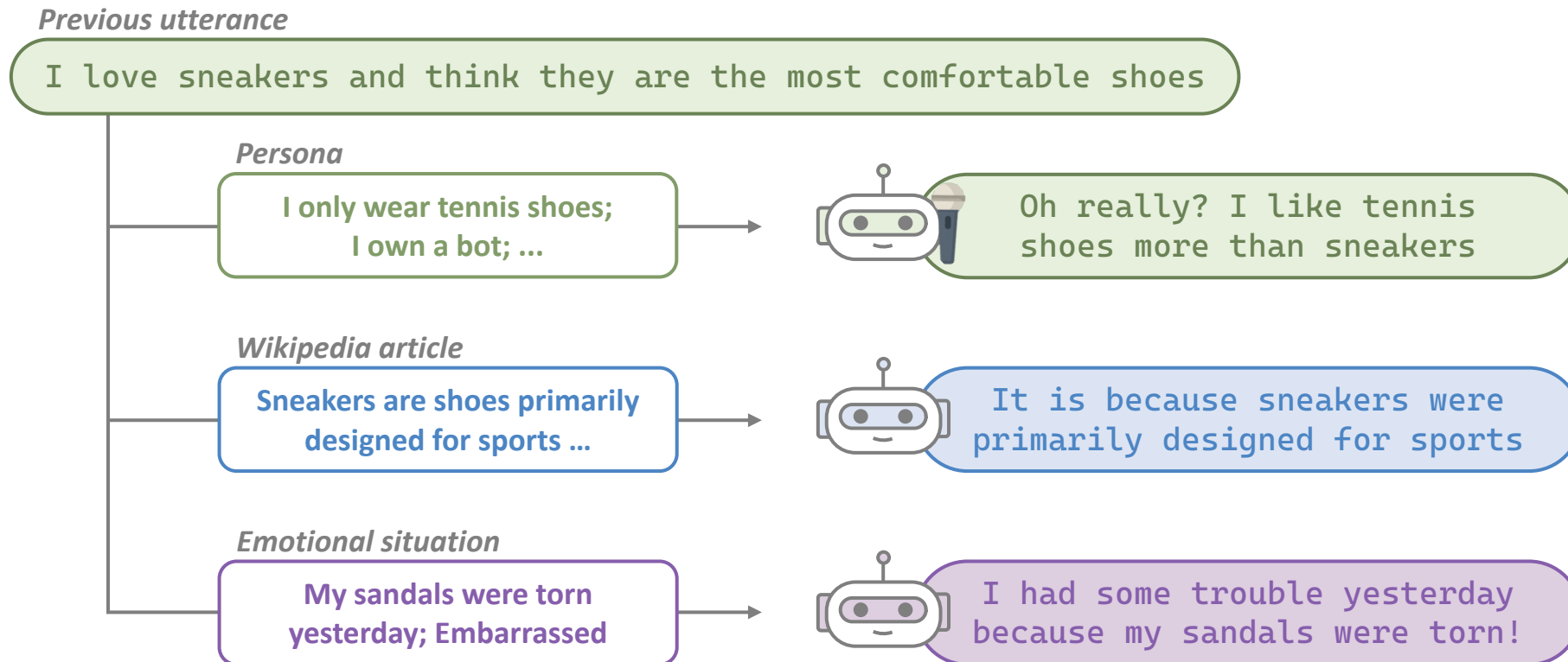
BotsTalk Framework

Phase 1: Simulate what to speak



BotsTalk Framework

Phase 1: Simulate what to speak



BotsTalk Framework

Phase 2: Check dialogue consistency

Persona

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Wikipedia article

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

Emotional situation

My sandals were torn
yesterday; Embarrassed

I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Persona

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Wikipedia article

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

Emotional situation

My sandals were torn
yesterday; Embarrassed

I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Challenge: *The generated response is likely to be semantically in conflict with other skill contexts*

Persona

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Wikipedia article

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

Emotional situation

My sandals were torn
yesterday; Embarrassed

I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Challenge: The generated response is likely to be semantically in conflict with other skill contexts



Why?

Persona

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Wikipedia article

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

Emotional situation

My sandals were torn
yesterday; Embarrassed

I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Challenge: *The generated response is likely to be semantically in conflict with other skill contexts*

➡ *The skill agents are not allowed to read other skill contexts of different skills*

Why?

Persona

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Wikipedia article

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

Emotional situation

My sandals were torn
yesterday; Embarrassed

I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI



A set of
skill contexts

I only wear tennis shoes;
I own a bot; ...

Sneakers are shoes primarily
designed for sports ...

My sandals were torn
yesterday; Embarrassed

Oh really? I like tennis
shoes more than sneakers

It is because sneakers were
primarily designed for sports

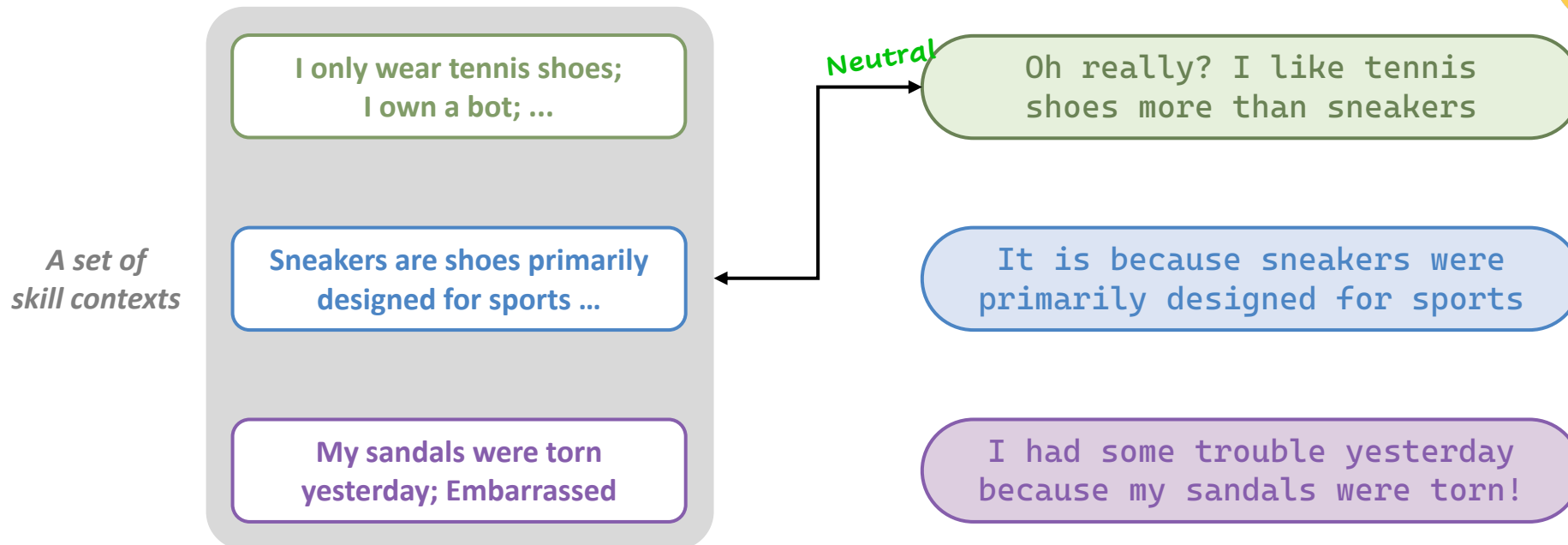
I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI

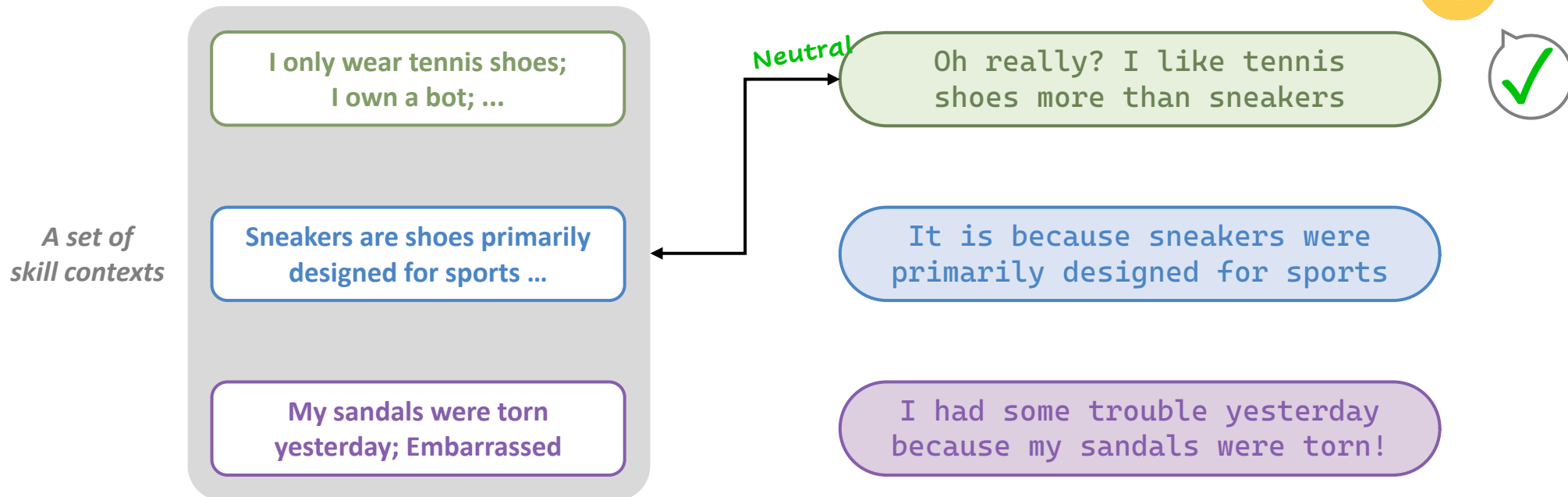


BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI

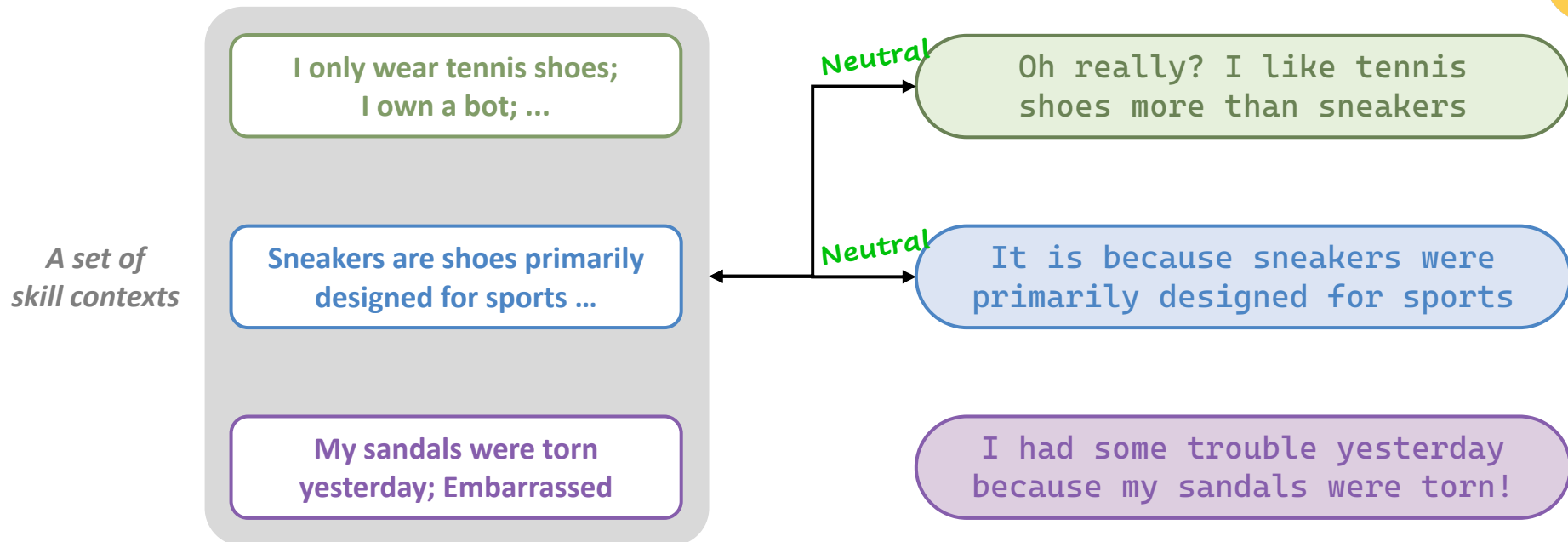


BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI



BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI



A set of
skill contexts

I only wear tennis shoes;
I own a bot; ...

Sneakers are shoes primarily
designed for sports ...

My sandals were torn
yesterday; Embarrassed

Neutral

Oh really? I like tennis
shoes more than sneakers

Neutral

It is because sneakers were
primarily designed for sports

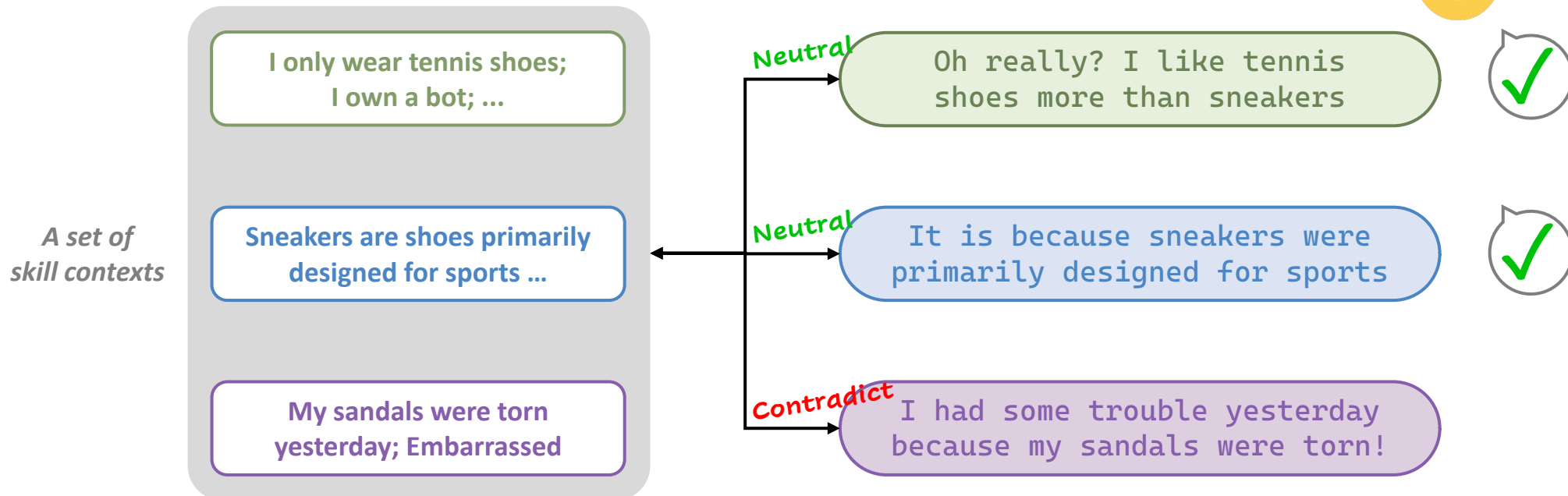
I had some trouble yesterday
because my sandals were torn!

BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI

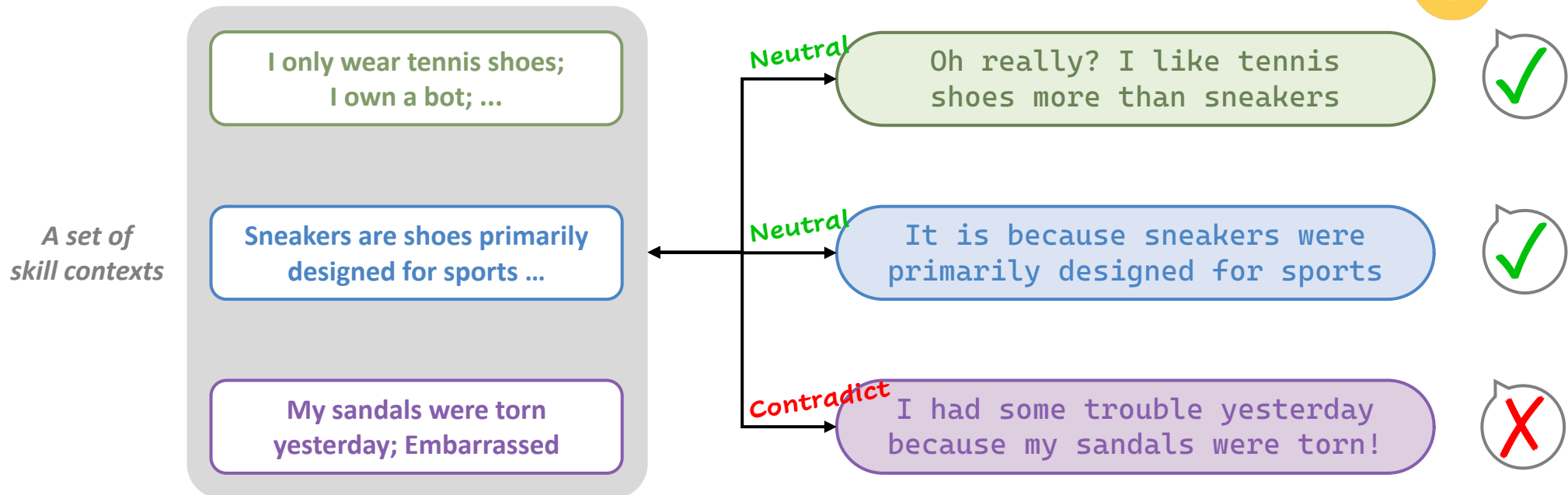


BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI



BotsTalk Framework

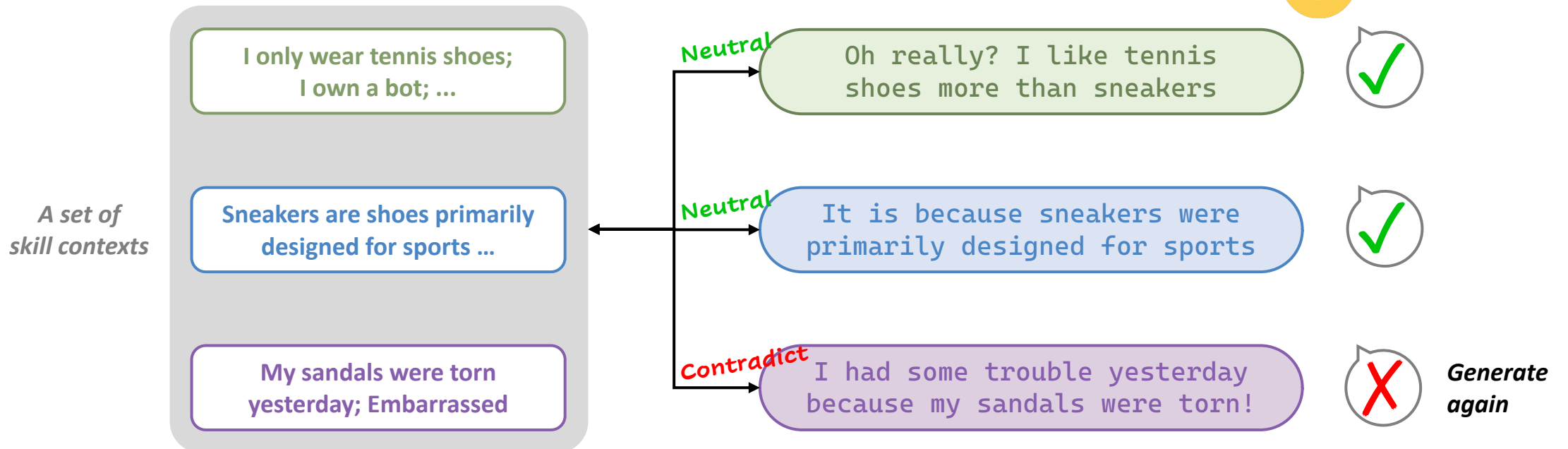
Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*

NLI



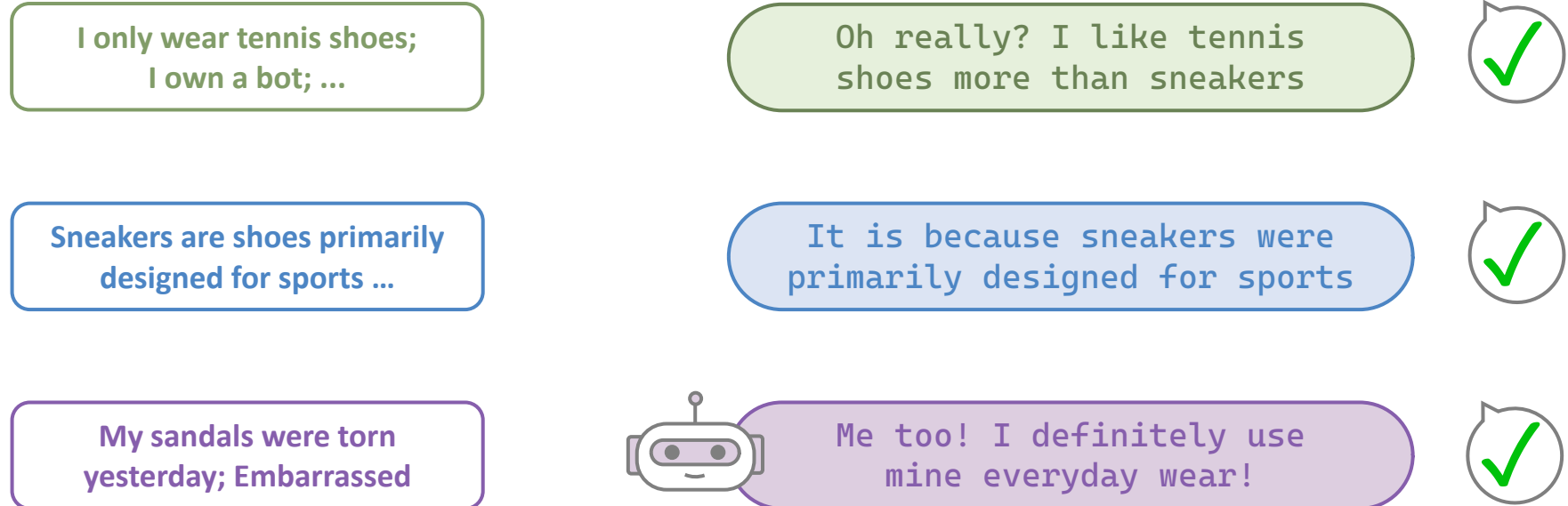
Generate again



BotsTalk Framework

Phase 2: Check dialogue consistency

Solution: *The moderator agent, who has access to all skill contexts, filters out conflicting response candidates for dialogue consistency*



BotsTalk Framework

Phase 3: Speak or pass the mic

Challenge: How to identify the most appropriate response to the dialogue context?

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Sneakers are shoes primarily
designed for sports ...

It is because sneakers were
primarily designed for sports

My sandals were torn
yesterday; Embarrassed

Me too! I definitely use
mine everyday wear!

BotsTalk Framework

Phase 3: Speak or pass the mic

Solution: *The active agent, who has a priority over the current dialogue flow, scores and ranks a set of response candidates*

I only wear tennis shoes;
I own a bot; ...

Oh really? I like tennis
shoes more than sneakers

Sneakers are shoes primarily
designed for sports ...

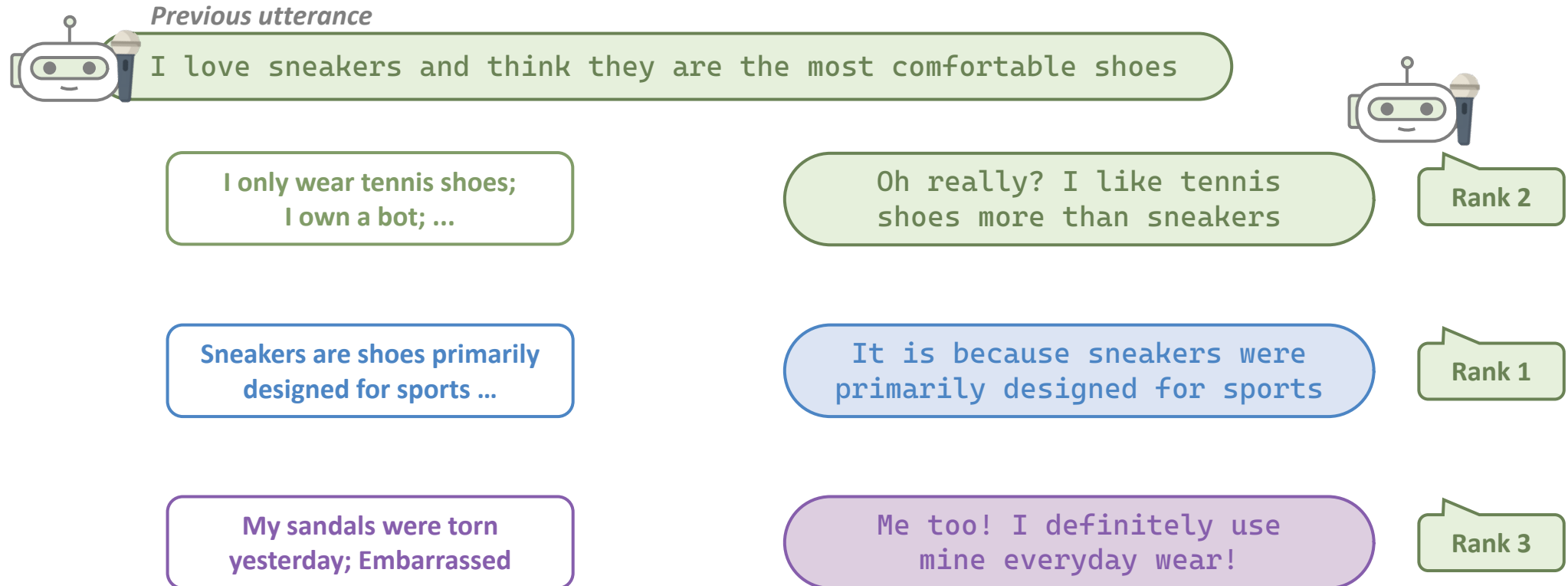
It is because sneakers were
primarily designed for sports

My sandals were torn
yesterday; Embarrassed

Me too! I definitely use
mine everyday wear!

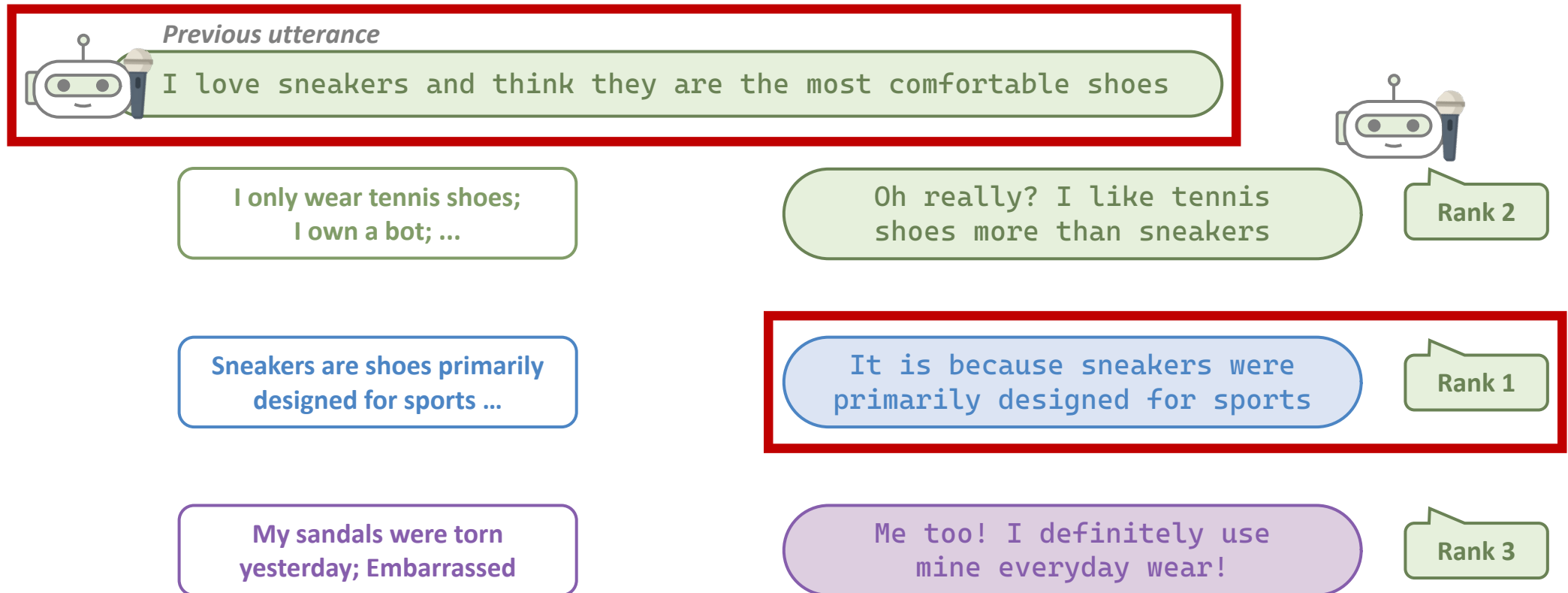
BotsTalk Framework

Phase 3: Speak or pass the mic



BotsTalk Framework

Phase 3: Speak or pass the mic



BotsTalk Framework

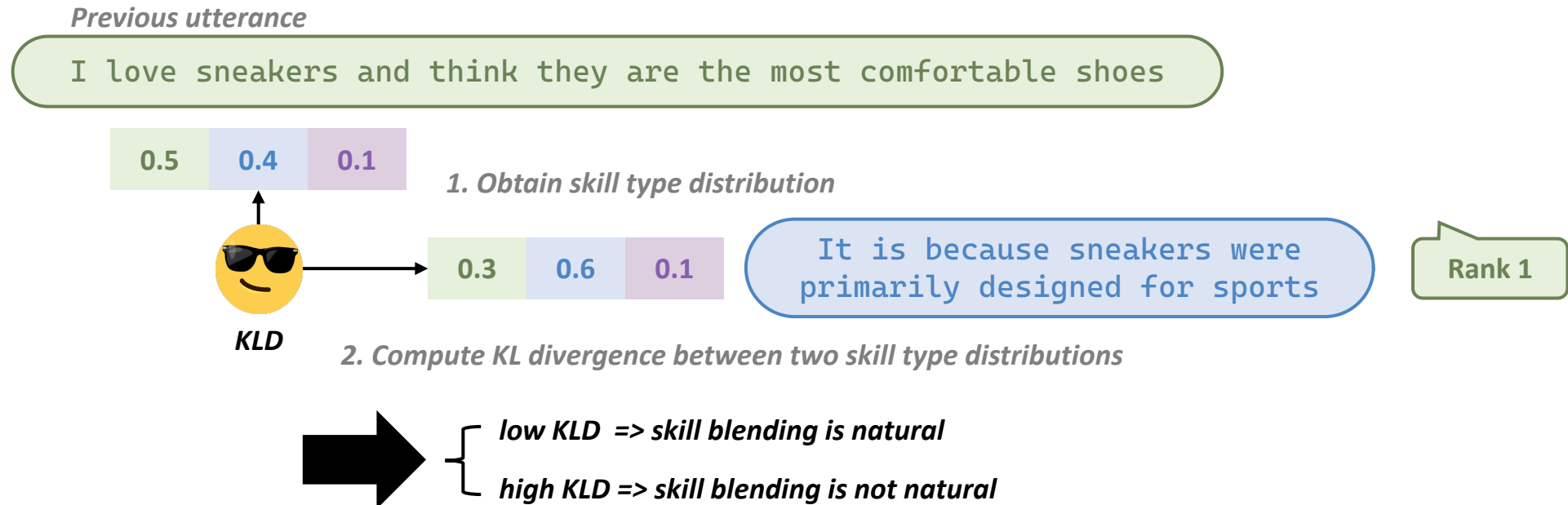
Phase 3: Speak or pass the mic



BotsTalk Framework

Phase 3: Speak or pass the mic

Solution: *For natural dialogue flow, the moderator agent compute KL divergence between skill distributions of previous utterance and rank 1 candidate*



BotsTalk Framework

Phase 3: Speak or pass the mic



Should I pass the mic?

If skill blending is natural (a.k.a. low KLD)

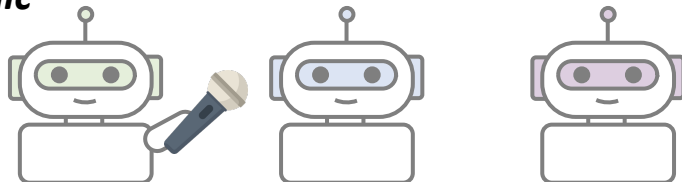


KLD

Rank 1

It is because sneakers were primarily designed for sports

Pass the mic



If skill blending is not natural (a.k.a. high KLD)

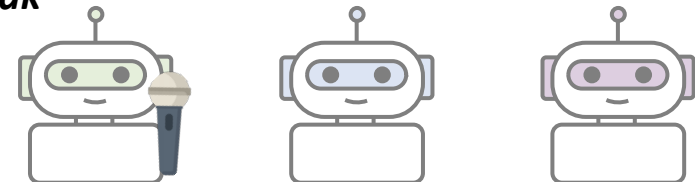


KLD

Rank 2

Oh really? I like tennis shoes more than sneakers

Speak



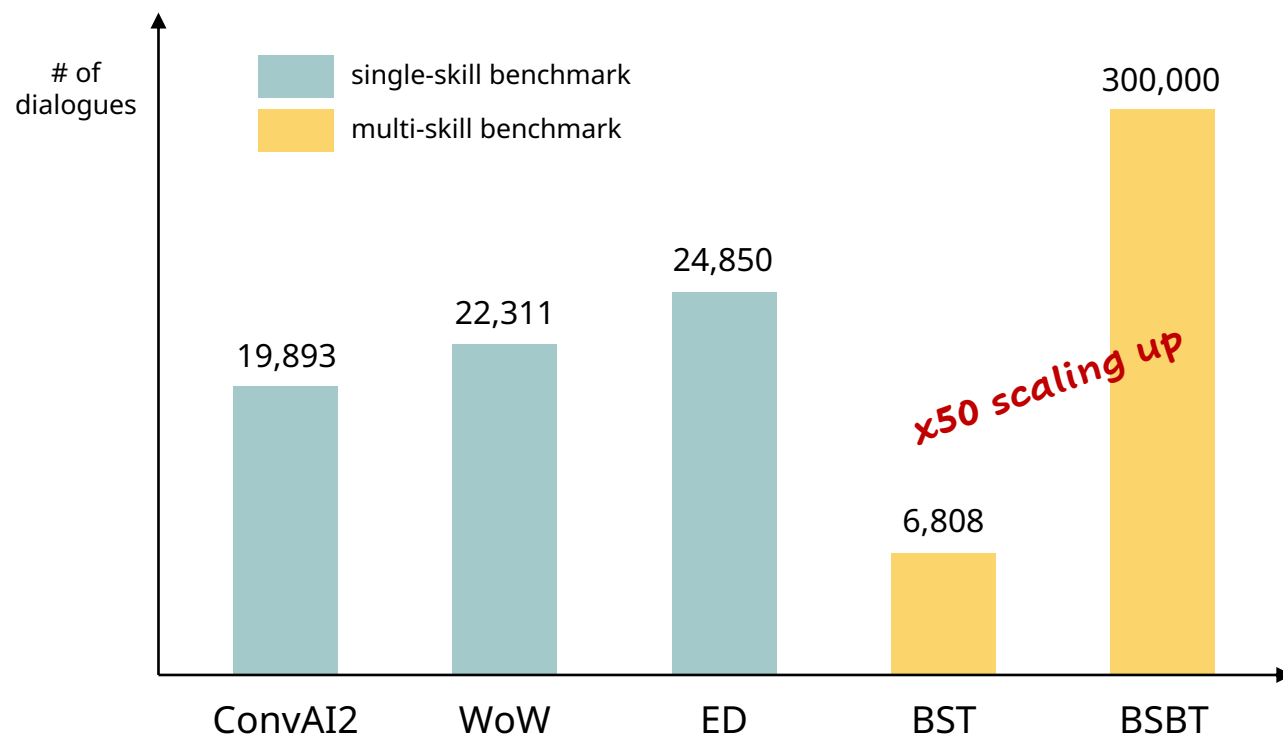
Skill context from ConvAI2 I like to play soccer; I like to read; ... Skill context from WoW Nike Inc. Skill context from ED I really like this girl at my job, but I am ... ; Apprehensive	Skill context from ConvAI2 I have 3 children; I am a karate black belt; ... Skill context from WoW Nike Inc.; ... multinational corporation ... Air Jordan ... Skill context from ED None
Dialogue A: Do you have much experience using the different types of cleats? Which do you like best? (P) B: I have a little. I also know about Air jordans, a brand of footwear also popular with athletes. (K) A: I enjoy Air jordans as well. I like to play soccer, and it's extremely hard to get good footwear. (P) B: I agree. Air jordans are good for calves, and it's really easy to pull off. (K) A: Air jordans are generally made with the most material so it makes sense they'd be easy to use. (K) B: And now shoes can go together with clothes as well, like any other type of material. (K) A: Do you know much about aeros then? They have excellent fit and beauty. (K) B: I have a hard time finding it but they are great shoes. (P) A: I hope you can find ones that are comfortable to you. (E) B: Yes. I hope you can get those shoes too. (E)	

Blended Skill BotsTalk (BSBT)

Table 2: A conversation from the BSBT dataset. Speaker A is given five personas, one topic, and a situation with an emotion (top left), while speaker B is given five personas, one topic, and seven knowledge resources (top right).

Blended Skill BotsTalk (BSBT)

Data statistics



- BotsTalk framework enables scaling up the data size to 50 times more than BST.

Blended Skill BotsTalk (BSBT)

Data quality

▼ Human evaluation for pairwise comparison between BST and BSBT datasets.

	Engagingness	Interestingness	Humanness
BST	43	47	44
BSBT	57	53	56

Blended Skill BotsTalk (BSBT)

Data quality

▼ Human evaluation for pairwise comparison between BST and BSBT datasets.

	Engagingness	Interestingness	Naturalness
BST	43	47	44
BSBT	57	53	56

- ▶ For all metrics, machine-written BSBT dataset achieves higher win percentages over human-written BST dataset, which validates the quality of our BSBT dataset.
- ▶ Our BotsTalk framework can be an effective and efficient alternative to crowdsourcing when constructing multi-skill dialogue datasets.

Experiments

Automatic Evaluation on BST benchmark

- ❑ To test our BSBT dataset, we consider both retrieval task and generative task.
- ❑ Baselines
 - ❑ Retrieval task: pre-trained Poly-encoder (Humeau et al., 2020) with 256-million parameters
 - ❑ Generative task: pre-trained Bart-large (Lewis et al., 2020)
- ❑ Metrics
 - ❑ Retrieval task: Recall@k (R@k), Mean Reciprocal Rank (MRR)
 - ❑ Generative task: BLEU-1, BLEU-2, BLEU-3, BLEU-4

Experiments

Automatic Evaluation on BST benchmark

- ❑ To test our BSBT dataset, we consider both retrieval task and generative task.

	Retrieval task (Poly-encoder (Humeau et al., 2020))				Generative (Bart-large (Lewis et al., 2020))			
	R@1	R@5	R@10	MRR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
BST	75.92	94.76	97.83	84.14	12.19	3.65	1.06	0.37
BSBT	80.68	95.79	98.16	87.39	11.92	3.74	1.28	0.57

Experiments

Automatic Evaluation on BST benchmark

- ❑ To test our BSBT dataset, we consider both retrieval task and generative task.

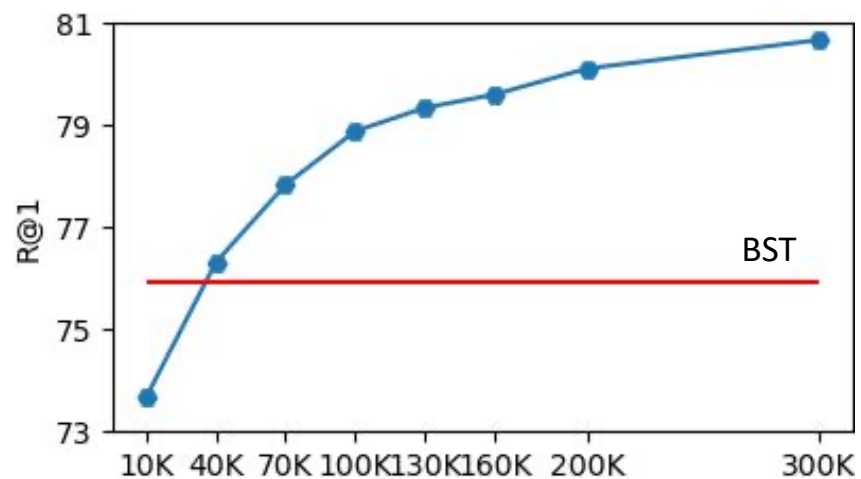
	Retrieval task (Poly-encoder (Humeau et al., 2020))				Generative (Bart-large (Lewis et al., 2020))			
	R@1	R@5	R@10	MRR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
BST	75.92	94.76	97.83	84.14	12.19	3.65	1.06	0.37
BSBT	80.68	95.79	98.16	87.39	11.92	3.74	1.28	0.57

- ▶ BSBT model outperforms BST model on most automatic metrics.
- ▶ Our BSBT dataset works properly as training resource to learn the ability of blending skills as well as grounding to various skills.

Experiments

Analysis on Scalability

- To explore the impact of BSBT size on the model performance, we fine-tune the retrieval architecture on the BSBT datasets of varying scales.

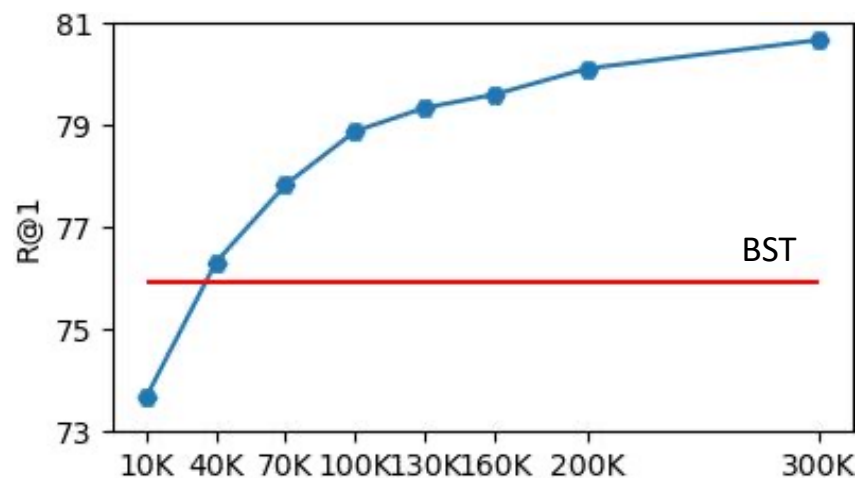


- ▲ The effect on the model performance by varying the number of dialogues in training set.

Experiments

Analysis on Scalability

- To explore the impact of BSBT size on the model performance, we fine-tune the retrieval architecture on the BSBT datasets of varying scales.



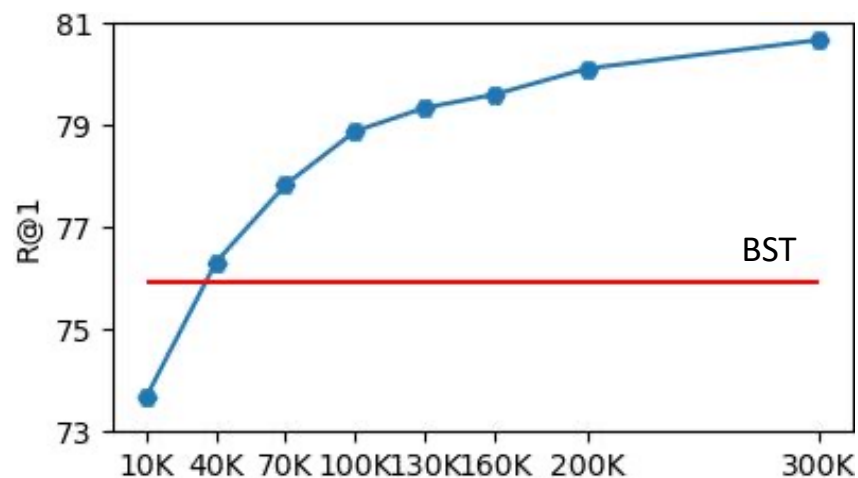
- BSBT300K model achieves a significant performance boost from BSBT10K model.

- ▲ The effect on the model performance by varying the number of dialogues in training set.

Experiments

Analysis on Scalability

- To explore the impact of BSBT size on the model performance, we fine-tune the retrieval architecture on the BSBT datasets of varying scales.



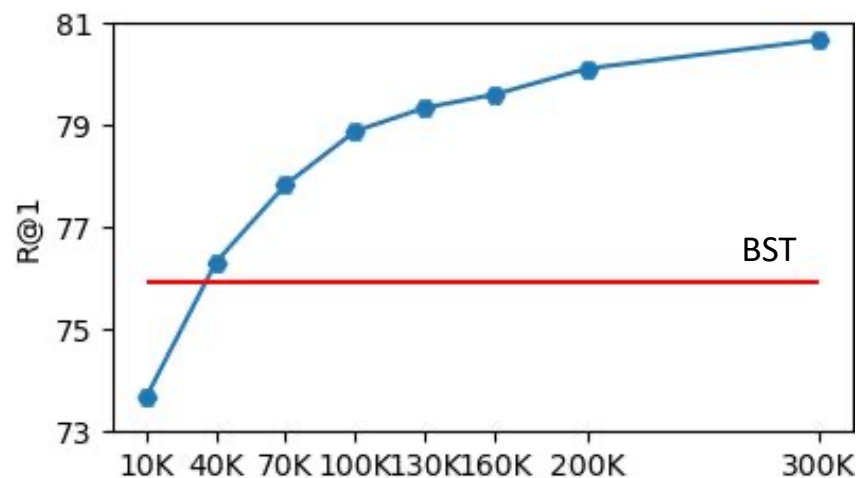
- ▲ The effect on the model performance by varying the number of dialogues in training set.

- ▶ BSBT300K model achieves a significant performance boost from BSBT10K model.
- ▶ Large-scale training is important.

Experiments

Analysis on Scalability

- To explore the impact of BSBT size on the model performance, we fine-tune the retrieval architecture on the BSBT datasets of varying scales.



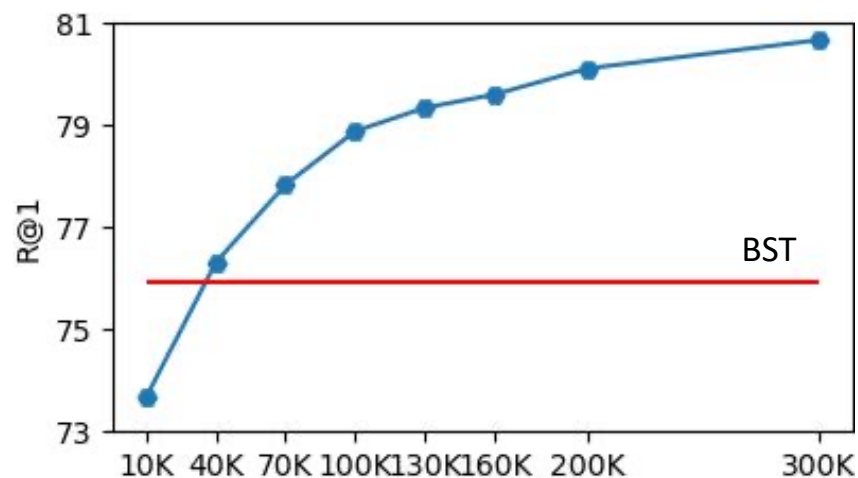
- ▲ The effect on the model performance by varying the number of dialogues in training set.

- ▶ BSBT300K model achieves a significant performance boost from BSBT10K model.
- ▶ Large-scale training is important.
- ▶ This reveals the potential of BSBT, as our dataset is collected by automatic framework without human intervention.

Experiments

Analysis on Scalability

- To explore the impact of BSBT size on the model performance, we fine-tune the retrieval architecture on the BSBT datasets of varying scales.



▲ The effect on the model performance by varying the number of dialogues in training set.

- ▶ BSBT300K model achieves a significant performance boost from BSBT10K model.
- ▶ Large-scale training is important.
- ▶ This reveals the potential of BSBT, as our dataset is collected by automatic framework without human intervention.
- ▶ Our BotsTalk framework is scalable with respect to the data size and skill type.

Experiments

Analysis on multi-task learning

- ❑ Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021).
- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

Experiments

Analysis on multi-task learning

- ❑ Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021).
- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

Experiments

Analysis on multi-task learning

- ❑ Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021).
- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

► MTL model lags behind BSBT model on BST benchmark.

Experiments

Analysis on multi-task learning

- ❑ Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021).
- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

- ▶ MTL model lags behind BSBT model on BST benchmark.
- ▶ MTL model performs noticeably better when fine-tuned on BSBT datasets in addition.

Experiments

Analysis on multi-task learning

- ❑ Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021).
- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

- ▶ MTL model lags behind BSBT model on BST benchmark.
- ▶ MTL model performs noticeably better when fine-tuned on BSBT datasets in addition.
- ▶ Such improvement indicates BSBT is orthogonally applicable to multi-tasking strategy.

Experiments

Analysis on multi-task learning

- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

The performance gain becomes marginal when the size of the dataset increases.

Experiments

Analysis on multi-task learning

- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

Why? 

The performance gain becomes marginal when the size of the dataset increases.

Experiments

Analysis on multi-task learning

- ❑ Therefore, we consider MTL model, a poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, Wizard of Wikipedia, and Empathetic Dialogues.
- ❑ To probe the effectiveness of BSBT as a training resource for multi-task training scheme, we further fine-tune the MTL model on BSBT datasets of varying scales sequentially.

	R@1	R@5	MRR
BST	75.92	94.76	84.14
BSBT	80.68	95.79	87.39
MTL	78.95	95.37	86.23
+ BSBT100K	80.94	95.79	86.92
+ BSBT200K	82.01	96.37	87.83
+ BSBT300K	82.10	96.79	88.04

- As multi-task learning and BSBT are parameterized and materialized knowledge for multi-skill dialogues respectively, there can be an overlap between the knowledge dialogue systems learn.

Why? 

The performance gain becomes marginal when the size of the dataset increases.

Summary

- ✓ We propose an automatic data construction framework **BotsTalk** for developing open-domain chatbots that are able to use diverse communicative skills.
- ✓ We further present **Blended Skill BotsTalk (BSBT)**, a large-scale multi-skill dialogue dataset comprising 300K conversations.
- ✓ We demonstrate that our framework can be an effective and efficient alternative to crowdsourcing.
- ✓ We demonstrate that our dataset is helpful for multi-skill dialogue systems which require an understanding of skill blending as well as skill grounding.
- ✓ Future work could extend the scope of skill types and tasks.



EMNLP
2022



Thank you for listening

Paper



Code/data



Contact: minnju@yonsei.ac.kr / cheris8@yonsei.ac.kr