

Leveraging Large Language Models for Automated Dialogue Analysis

Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi

SIGDIAL 2023



EMORY
UNIVERSITY



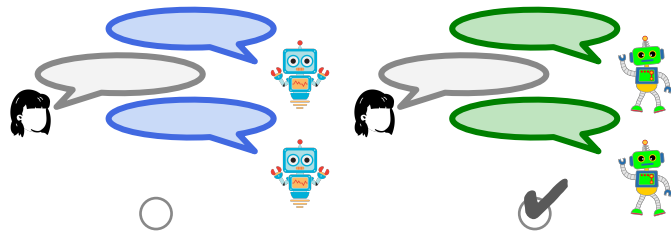


Pairwise Selections

Dialogue Ratings



Pairwise Selections

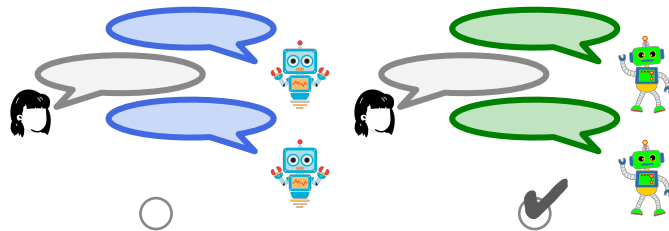


Quality

Dialogue Ratings

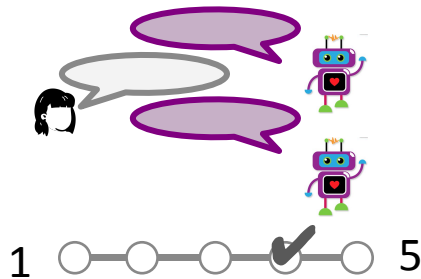


Pairwise Selections



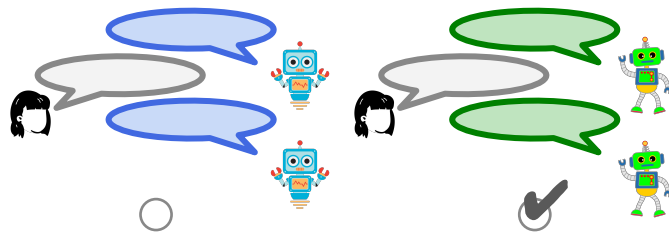
Quality

Dialogue Ratings

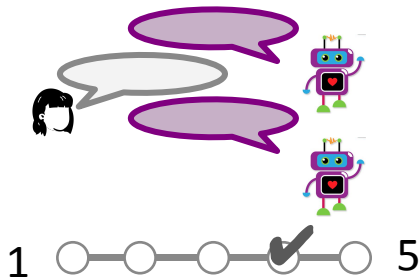




Pairwise Selections



Dialogue Ratings



Quality

- Consistency
- Emotional Understanding
- Engagingness
- Grammaticality
- Informativeness
- Proactivity
- Relevance

My favorite movie is the Lion King.



I like it too! Do you want to watch it with me?



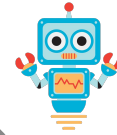
No, I don't like that movie.



I'd love to. Do you like it?



My favorite movie is the Lion King.



I like it too! Do you want to watch it with me?



No, I don't like that movie.



I'd love to. Do you like it?



Both Inconsistent!

My favorite movie is the Lion King.



I like it too! Do you want to watch it with me?



Self
Contradiction

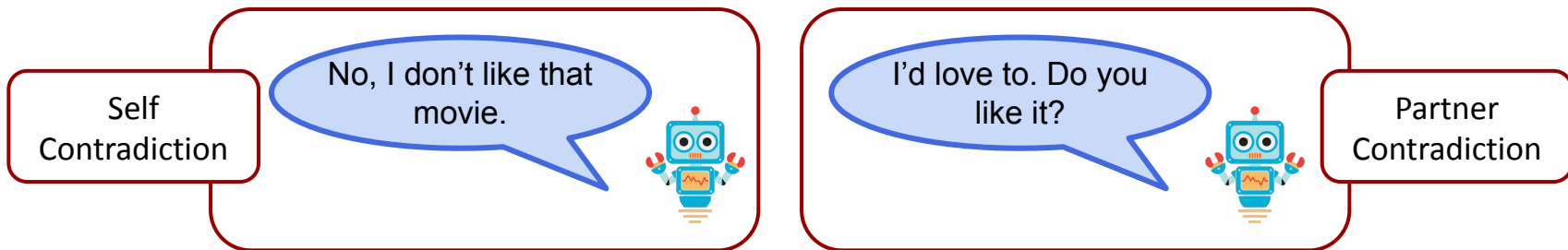
No, I don't like that movie.



I'd love to. Do you like it?



Both Inconsistent!



Both Inconsistent!

Behavior-Specific Classifiers

EPITOME

Sharma et al. A computational approach to understanding empathy expressed in text-based mental health support. *EMNLP 2020*.

Empathetic Response

Lack of Empathy

Behavior-Specific Classifiers

EPITOME

Sharma et al. A computational approach to understanding empathy expressed in text-based mental health support. *EMNLP 2020*.

Empathetic Response

Lack of Empathy

DECODE

Nie et al. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. *ACL 2021*.

Self Contradiction

Behavior-Specific Classifiers

EPITOME

Sharma et al. A computational approach to understanding empathy expressed in text-based mental health support. *EMNLP 2020*.

Empathetic Response

Lack of Empathy

DECODE

Nie et al. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. *ACL 2021*.

Self Contradiction

DialFact / FaithDial

Gupta et al. Dialfact: A benchmark for fact-checking in dialogue. *ACL 2022*.

Dziri et al. Faithdial: A faithful benchmark for information-seeking dialogue. *TACL 2022*.

Incorrect Fact

Behavior-Specific Classifiers

EPITOME

Sharma et al. A computational approach to understanding empathy expressed in text-based mental health support. *EMNLP 2020*.

Empathetic Response

Lack of Empathy

DECODE

Nie et al. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. *ACL 2021*.

Self Contradiction

DialFact / FaithDial

Gupta et al. Dialfact: A benchmark for fact-checking in dialogue. *ACL 2022*.

Dziri et al. Faithdial: A faithful benchmark for information-seeking dialogue. *TACL 2022*.

Incorrect Fact

Dialogue Breakdown

Lin and Ng. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues. *AAAI 2022*.

Ignoring Partner

Irrelevant

Behavior-Specific Classifiers

EPITOME

Sharma et al. A computational approach to understanding empathy expressed in text-based mental health support. *EMNLP 2020*.

Empathetic Response

Lack of Empathy

DECODE

Nie et al. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. *ACL 2021*.

Self Contradiction

DialFact / FaithDial

Gupta et al. Dialfact: A benchmark for fact-checking in dialogue. *ACL 2022*.

Dziri et al. Faithdial: A faithful benchmark for information-seeking dialogue. *TACL 2022*.

Incorrect Fact

Dialogue Breakdown

Lin and Ng. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues. *AAAI 2022*.

Ignoring Partner

Irrelevant

Redundant

Partner Contradiction

Commonsense Violation

Can the detection of *any* dialogue behavior be automated at low-cost?

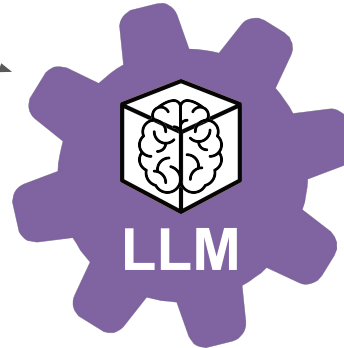


A response is repetitive if:

- it repeats something from earlier in the dialogue
- it includes asking a question whose answer has been already shared

If any part of the response is repetitive, then it should be labeled as repetitive.

Note that sometimes repetition is useful, such as for emphasis, acknowledgement, clarification, or elaboration, and in these cases it should NOT be labeled as repetitive.

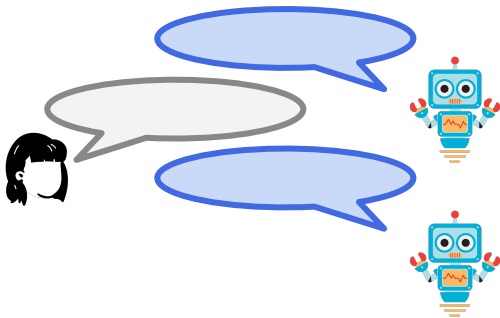
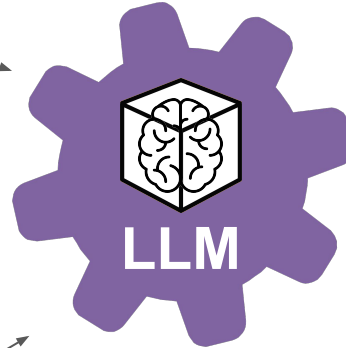


A response is repetitive if:

- it repeats something from earlier in the dialogue
- it includes asking a question whose answer has been already shared

If any part of the response is repetitive, then it should be labeled as repetitive.

Note that sometimes repetition is useful, such as for emphasis, acknowledgement, clarification, or elaboration, and in these cases it should NOT be labeled as repetitive.

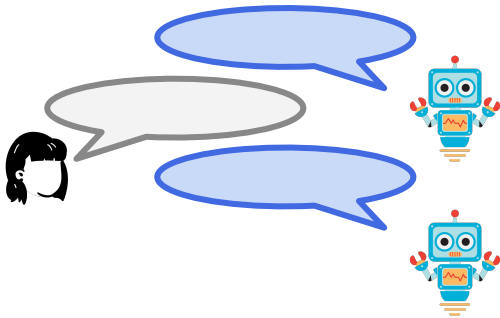
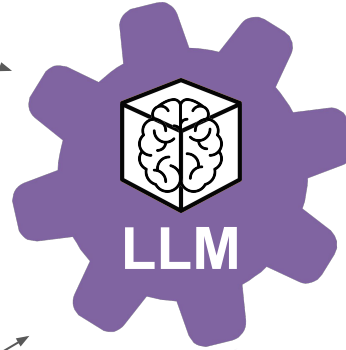


A response is repetitive if:

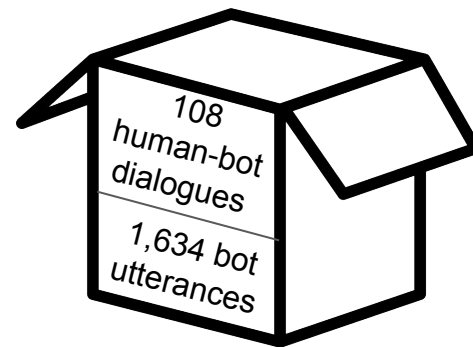
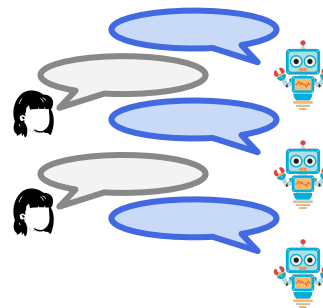
- it repeats something from earlier in the dialogue
- it includes asking a question whose answer has been already shared

If any part of the response is repetitive, then it should be labeled as repetitive.

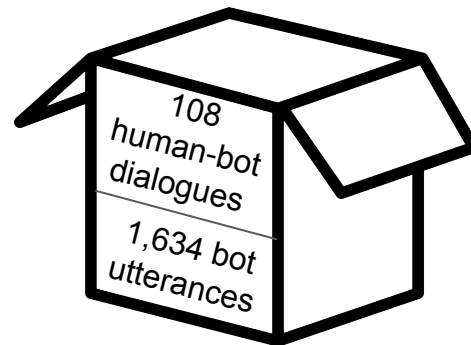
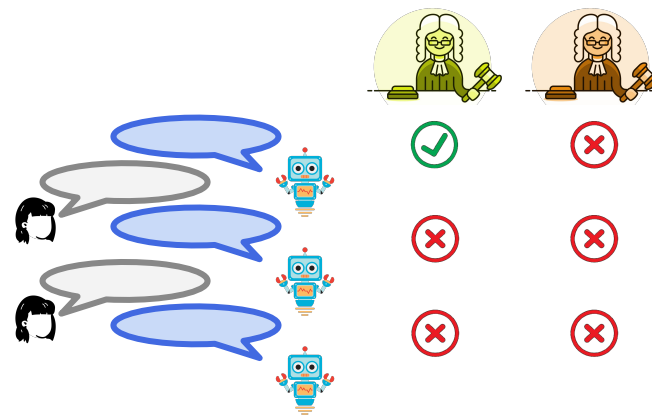
Note that sometimes repetition is useful, such as for emphasis, acknowledgement, clarification, or elaboration, and in these cases it should NOT be labeled as repetitive.



ABC-Eval Dataset



ABC-Eval Dataset



ABC-Eval Dataset

Dialogue Behaviors

Ignoring Partner

Incorrect Fact

Irrelevant

Commonsense Violation

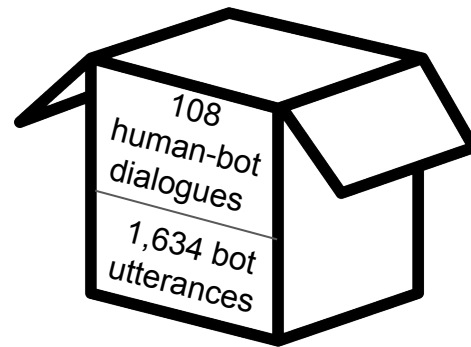
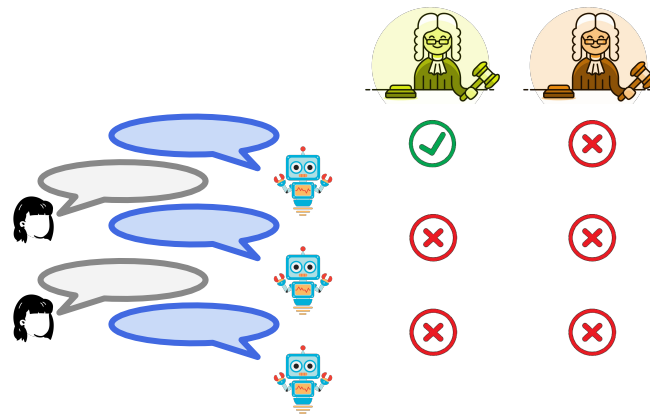
Redundant

Empathetic Response

Self Contradiction

Lack of Empathy

Partner Contradiction





Human Instructions

Empathy in Dialogue

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.



Human Instructions

Empathy in Dialogue

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

Prompt

Behavior Description

Behavior Question



Human Instructions

Empathy in Dialogue

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

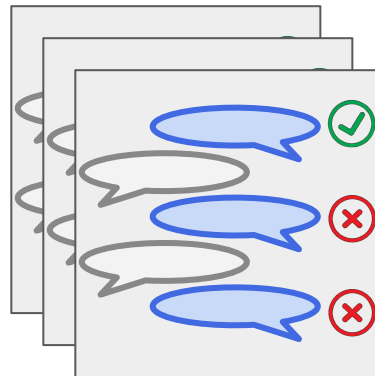
Prompt

Behavior Description

Behavior Question



Human Training Examples



Human Instructions

Empathy in Dialogue

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

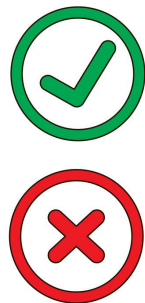
- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

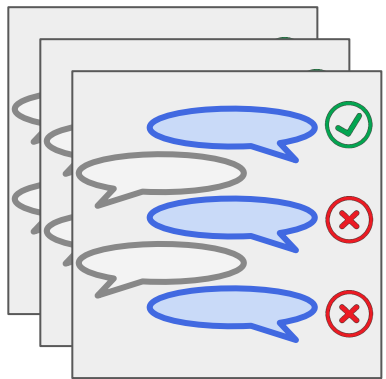
Prompt

Behavior Description

Behavior Question



Human Training Examples



Human Instructions

Empathy in Dialogue

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

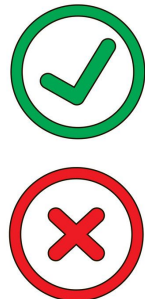
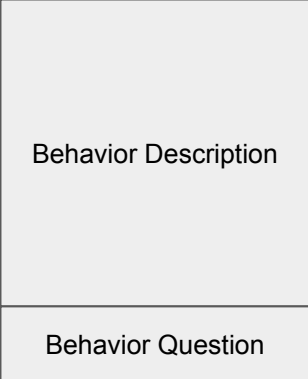
- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

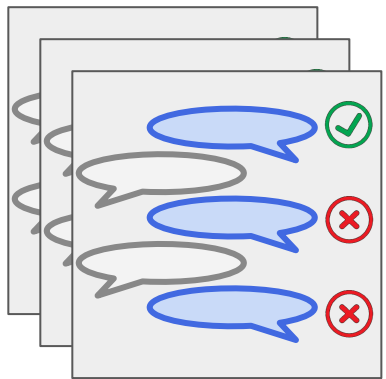
- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

Prompt



Human Training Examples



Empathetic

1. Speaker 1: Yes, I told you so.
2. Speaker 2: Cool! Personally, I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?
3. Speaker 1: More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

Is this an empathetic response by Speaker 2:

Speaker 2: I understand, it can definitely be stressful. What would you change about your school, if you could?

A response is empathetic when Speaker 2 does ONE of the following:

- clearly demonstrates an understanding of Speaker 1's emotions
- reacts with the appropriate sentiment or emotion to Speaker 1's shared experience
- understands or appropriately reacts to Speaker 1's experience or emotions
- appropriately reassures, encourages, or supports Speaker 1

Commonsense Violation

1. Speaker 1: Do you know any other states beside Florida?
2. Speaker 2: Yes, there are a few other states that border florida, such as texas, louisiana, mississippi, alabama, georgia, and tennessee.
3. Speaker 1: What state are you from?

If this were the next response in the dialogue, would it contradict commonsense:

Speaker 2: I'm from the state of new york. I was born and raised there, but my family moved to florida when I was a baby.

To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

Do NOT mark responses that don't make sense because they:

- are off-topic or irrelevant as responses
- don't have any clear meaning (e.g. overly vague or ill-formed responses)

Empathetic

1. Speaker 1: Yes, I told you so.
2. Speaker 2: Cool! Personally, I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?
3. Speaker 1: More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

Is this an empathetic response by Speaker 2:

Speaker 2: I understand, it can definitely be stressful. What would you change about your school, if you could?

A response is empathetic when Speaker 2 does ONE of the following:

- clearly demonstrates an understanding of Speaker 1's emotions
- reacts with the appropriate sentiment or emotion to Speaker 1's shared experience
- understands or appropriately reacts to Speaker 1's experience or emotions
- appropriately reassures, encourages, or supports Speaker 1

Commonsense Violation

1. Speaker 1: Do you know any other states beside Florida?
2. Speaker 2: Yes, there are a few other states that border florida, such as texas, louisiana, mississippi, alabama, georgia, and tennessee.
3. Speaker 1: What state are you from?

If this were the next response in the dialogue, would it contradict commonsense:

Speaker 2: I'm from the state of new york. I was born and raised there, but my family moved to florida when I was a baby.

To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

Do NOT mark responses that don't make sense because they:

- are off-topic or irrelevant as responses
- don't have any clear meaning (e.g. overly vague or ill-formed responses)

Empathetic

1. Speaker 1: Yes, I told you so.
2. Speaker 2: Cool! Personally, I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?
3. Speaker 1: More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

Is this an empathetic response by Speaker 2:

Speaker 2: I understand, it can definitely be stressful. What would you change about your school, if you could?

A response is empathetic when Speaker 2 does ONE of the following:

- clearly demonstrates an understanding of Speaker 1's emotions
- reacts with the appropriate sentiment or emotion to Speaker 1's shared experience
- understands or appropriately reacts to Speaker 1's experience or emotions
- appropriately reassures, encourages, or supports Speaker 1

Commonsense Violation

1. Speaker 1: Do you know any other states beside Florida?
2. Speaker 2: Yes, there are a few other states that border florida, such as texas, louisiana, mississippi, alabama, georgia, and tennessee.
3. Speaker 1: What state are you from?

If this were the next response in the dialogue, would it contradict commonsense:

Speaker 2: I'm from the state of new york. I was born and raised there, but my family moved to florida when I was a baby.

To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

Do NOT mark responses that don't make sense because they:

- are off-topic or irrelevant as responses
- don't have any clear meaning (e.g. overly vague or ill-formed responses)

Empathetic

1. Speaker 1: Yes, I told you so.
2. Speaker 2: Cool! Personally, I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?
3. Speaker 1: More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

Is this an empathetic response by Speaker 2:

Speaker 2: I understand, it can definitely be stressful. What would you change about your school, if you could?

A response is empathetic when Speaker 2 does ONE of the following:

- clearly demonstrates an understanding of Speaker 1's emotions
- reacts with the appropriate sentiment or emotion to Speaker 1's shared experience
- understands or appropriately reacts to Speaker 1's experience or emotions
- appropriately reassures, encourages, or supports Speaker 1

Commonsense Violation

1. Speaker 1: Do you know any other states beside Florida?
2. Speaker 2: Yes, there are a few other states that border florida, such as texas, louisiana, mississippi, alabama, georgia, and tennessee.
3. Speaker 1: What state are you from?

If this were the next response in the dialogue, would it contradict commonsense:

Speaker 2: I'm from the state of new york. I was born and raised there, but my family moved to florida when I was a baby.

To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

Do NOT mark responses that don't make sense because they:

- are off-topic or irrelevant as responses
- don't have any clear meaning (e.g. overly vague or ill-formed responses)

Empathetic

1. Speaker 1: Yes, I told you so.
2. Speaker 2: Cool! Personally, I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?
3. Speaker 1: More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

Is this an empathetic response by Speaker 2:

Speaker 2: I understand, it can definitely be stressful. What would you change about your school, if you could?

A response is empathetic when Speaker 2 does ONE of the following:

- clearly demonstrates an understanding of Speaker 1's emotions
- reacts with the appropriate sentiment or emotion to Speaker 1's shared experience
- understands or appropriately reacts to Speaker 1's experience or emotions
- appropriately reassures, encourages, or supports Speaker 1

Commonsense Violation

1. Speaker 1: Do you know any other states beside Florida?
2. Speaker 2: Yes, there are a few other states that border florida, such as texas, louisiana, mississippi, alabama, georgia, and tennessee.
3. Speaker 1: What state are you from?

If this were the next response in the dialogue, would it contradict commonsense:

Speaker 2: I'm from the state of new york. I was born and raised there, but my family moved to florida when I was a baby.

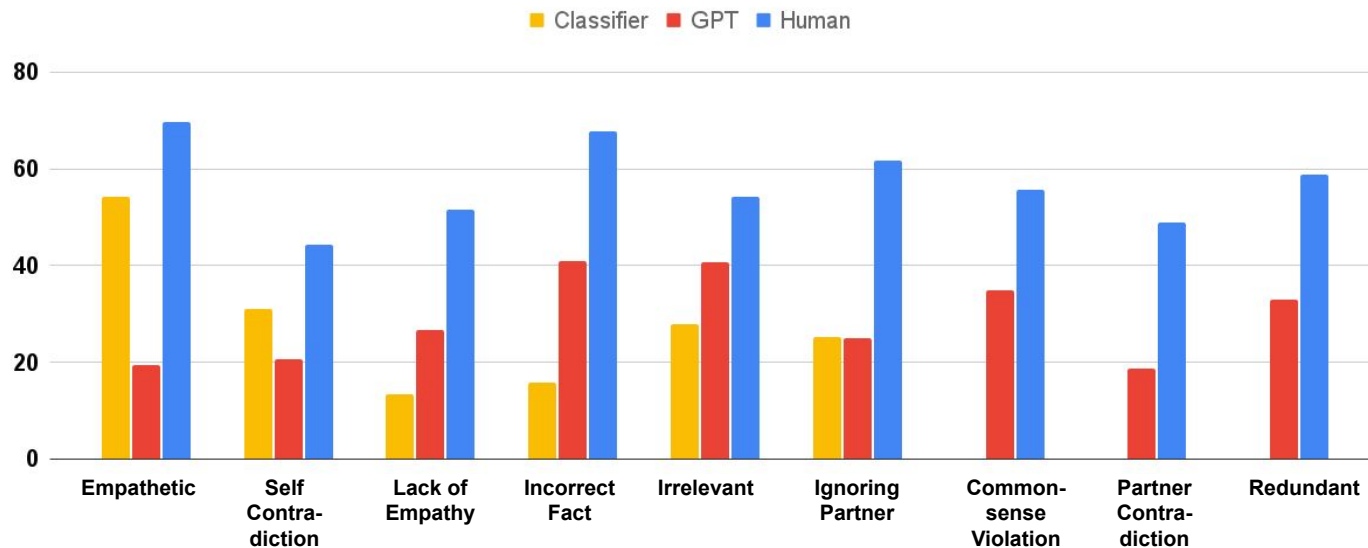
To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

Do NOT mark responses that don't make sense because they:

- are off-topic or irrelevant as responses
- don't have any clear meaning (e.g. overly vague or ill-formed responses)

F1 Score on Positive Examples of Dialogue Behaviors



Behavior-Specific Classifiers

EPITOME

Empathetic Response

Lack of Empathy

DECODE

Self Contradiction

DialFact / FaithDial

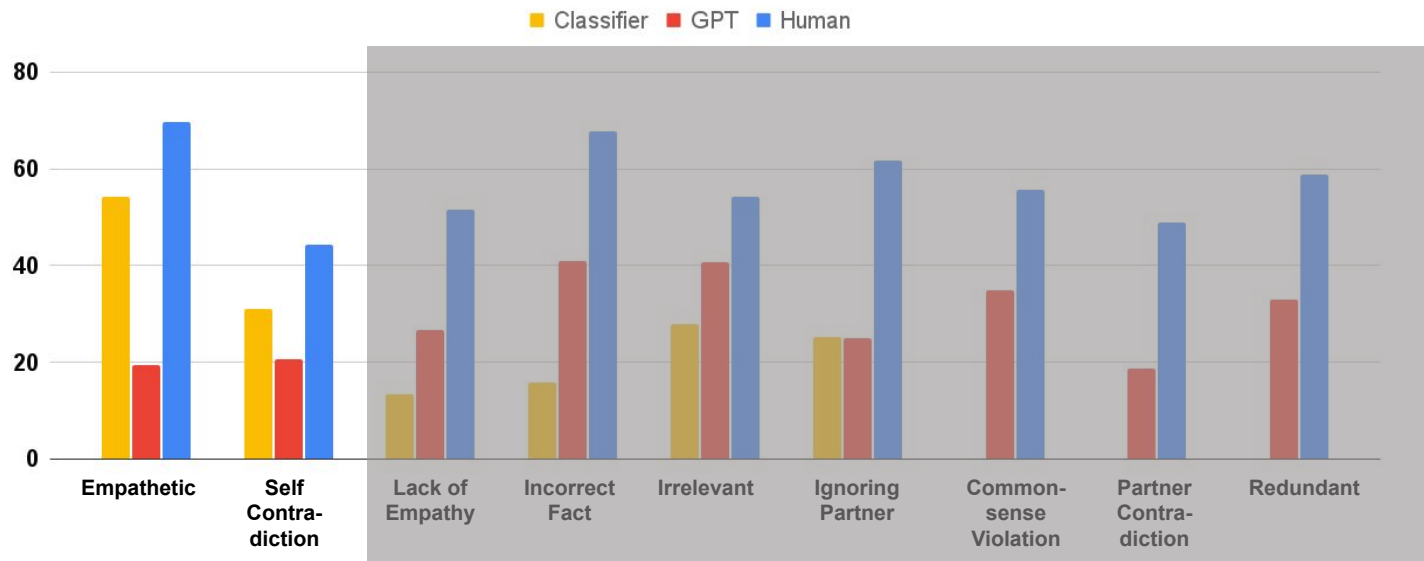
Incorrect Fact

Dialogue Breakdown

Ignoring Partner

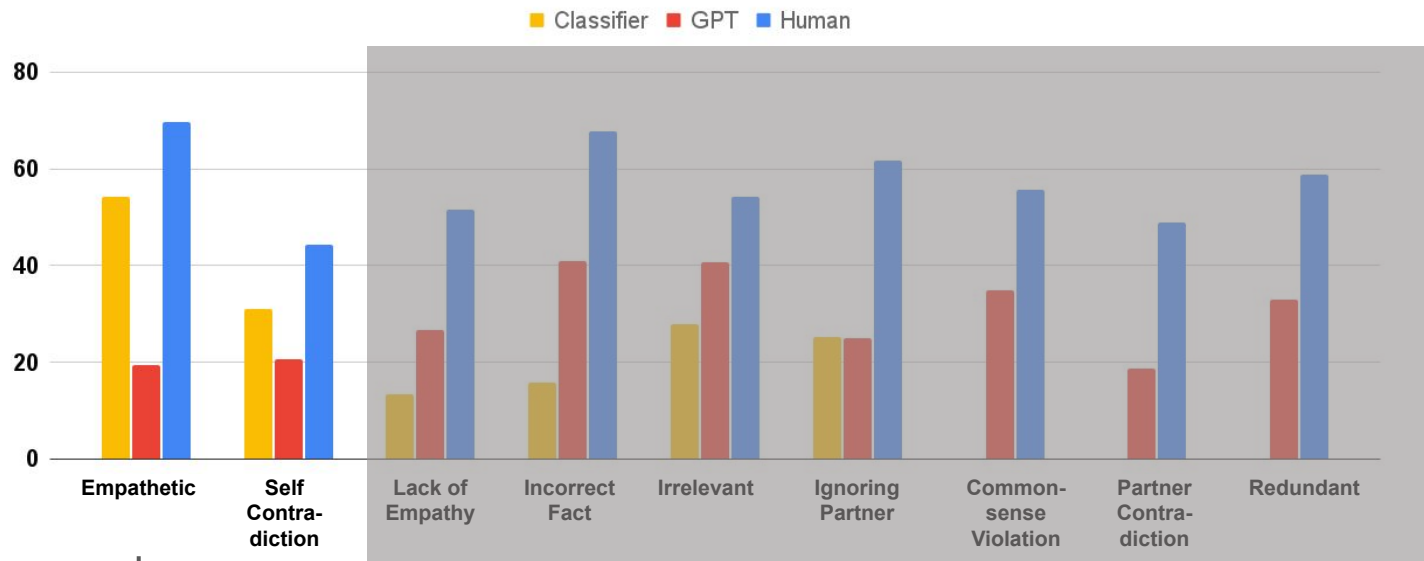
Irrelevant

F1 Score on Positive Examples of Dialogue Behaviors



Classifier outperforms GPT for minority of behaviors

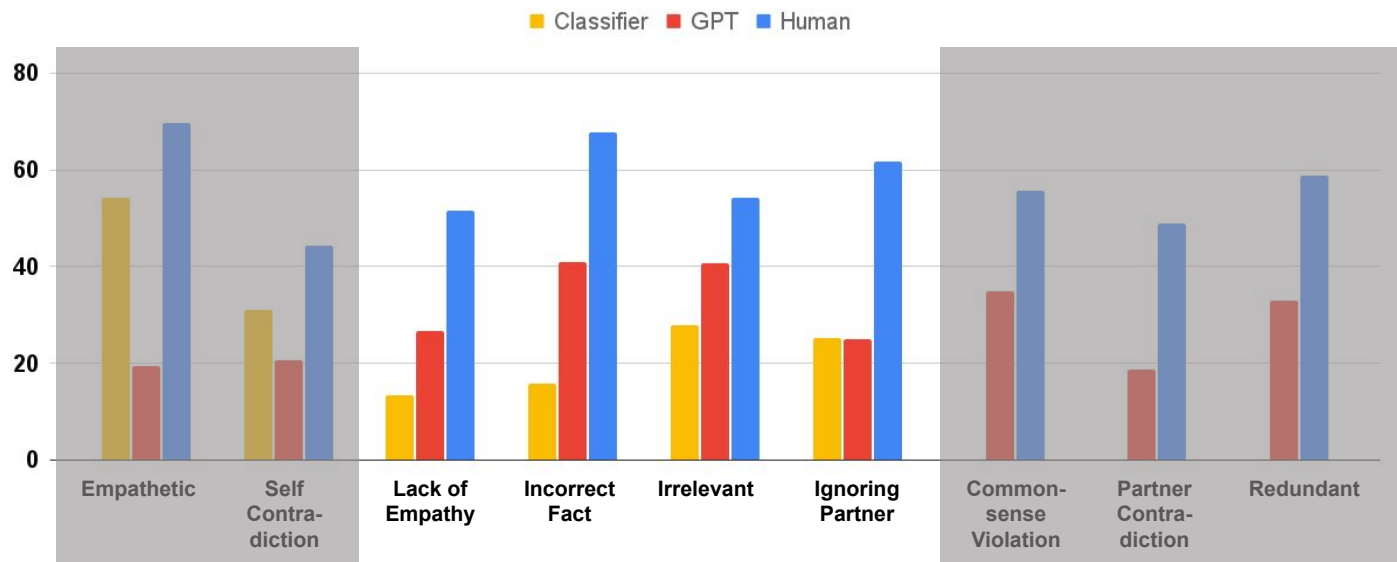
F1 Score on Positive Examples of Dialogue Behaviors



Classifier (EPITOME) predicts >80% of responses as empathetic, whereas Humans label <40% as empathetic

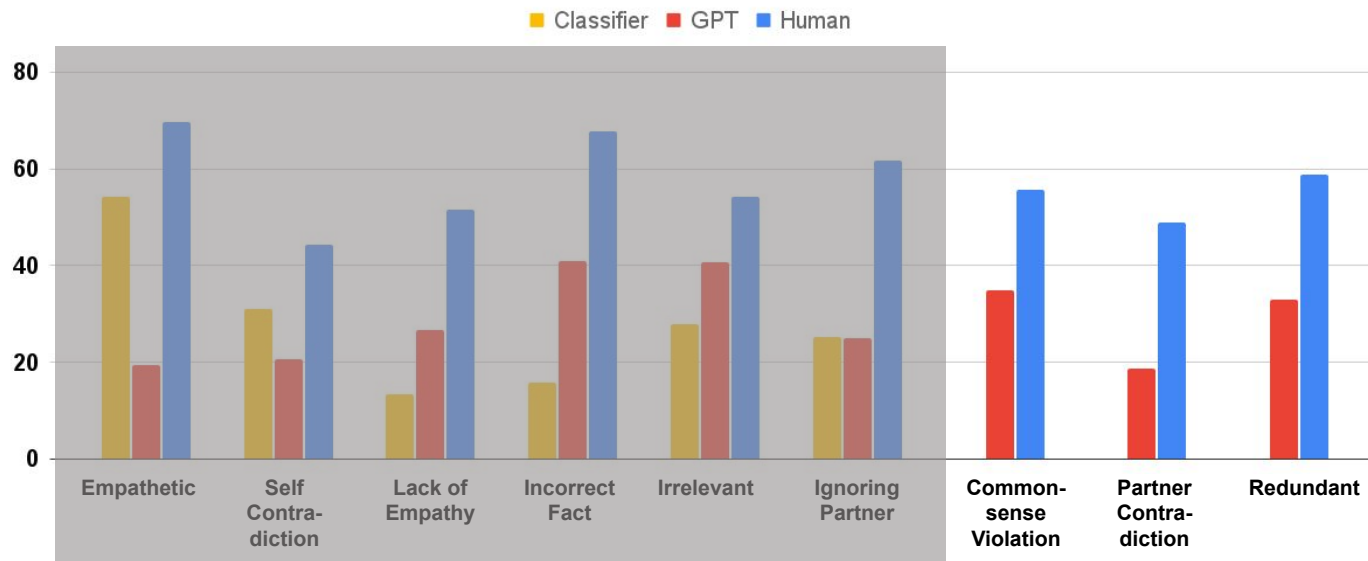
Low precision!

F1 Score on Positive Examples of Dialogue Behaviors



GPT more capable of detecting many behaviors than supervised classifiers!

F1 Score on Positive Examples of Dialogue Behaviors



GPT exhibits success even for behaviors currently without supervised classifiers!

GPT Reasoning

Speaker 1: Do you like historical dramas?

Speaker 2: Yes, I do! I love historical dramas! I especially like the ones set in the middle ages.

Speaker 1: Which is your favorite?

Speaker 2: My favorite is "the shawshank redemption" by martin scorsese. It's a classic!

Speaker 1: I love that movie. It is very interesting.

Speaker 2: That's true! I think it's interesting that the film was inspired by the real-life events that happened in the early 1900s.

Speaker 1: Have you ever been to prison?

Speaker 2: No, I've never been in prison. I've always wanted to go to prison, though.

Commonsense Violation

Reasoning: The response "I've always wanted to go to prison, though" contradicts common knowledge and is highly illogical. It is not a normal desire for someone to want to go to prison. Therefore, it would contradict commonsense.

Decision: [YES]

GPT Reasoning

Speaker 1: Do you like historical dramas?

Speaker 2: Yes, I do! I love historical dramas! I especially like the ones set in the middle ages.

Speaker 1: Which is your favorite?

Speaker 2: My favorite is "the shawshank redemption" by martin scorsese. It's a classic!

Speaker 1: I love that movie. It is very interesting.

Speaker 2: That's true! I think it's interesting that the film was inspired by the real-life events that happened in the early 1900s.

Speaker 1: Have you ever been to prison?

Speaker 2: No, I've never been in prison. I've always wanted to go to prison, though.

Commonsense Violation

Reasoning: The response "I've always wanted to go to prison, though" contradicts common knowledge and is highly illogical. It is not a normal desire for someone to want to go to prison. Therefore, it would contradict commonsense.

Decision: [YES]

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

- 40 failure examples from GPT for each dialogue behavior
- Categorized failure cause based on outputted reasoning

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

Lack of world and semantic knowledge is leading source of errors

Lack of Intelligence

(S1) I don't have a trip planned, but maybe I'll go somewhere soon

(S2) I'm sure you will! You'll love it! Have a wonderful trip! I'll see you soon!

[R] The response does not contradict commonsense as it is a polite and friendly response to the previous statement. It does not make any unfounded assumptions, is not illogical or self-contradictory, and does not ask a question where the answer is already obvious. **It is a common response to wish someone a good trip even if they have not planned one yet.**

Inexperience

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

Context management issues also contribute substantially to GPT failures

Context Management

(S1) I do! I have a younger brother. What about you?

(S2) You have a brother? Cool! Do you guys get along?

(S1) We used to fight a lot as kids, but we get along well now.

(S2) That's awesome. Personally, I always wanted an older sister so I could have someone to look up to. Are you older than your brother?

[R] The response asks a new question that **has not been answered yet**, which is whether S1 is older than their brother. **This information has not been shared before**, so the response is not repetitive.

History Forgetfulness

(S2) I love pop! What is your favorite R&B artist? Do you like Nicki Minaj?

(S1) I do like Nicki!

(S2) I like her too, she's so talented. Do either of your parents work? My mom is a teacher.

[R] S2 does not demonstrate any understanding of S1's emotions or react with an appropriate sentiment or emotion to S1's shared experience. **Instead, S2 changes the subject to ask about S1's parents.** This response is not empathetic.

Selective Attention

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

Tendency to stray from intended behavior definitions also hinders performance

Instruction Following

(S2) Cool! I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?

(S1) More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

(S2) I understand, it can definitely be stressful. What would you change about your school, if you could?

[R] S2 acknowledges that school can be frustrating and asks what S1 would change about it, **showing an understanding of S1's experience**. However, **they do not react with a specific sentiment or emotion, nor do they offer reassurance or support**.

Exhaustive

(S2) What is the name of this dorm? Do you have a roommate? How are you liking it?

(S1) It's MLAO. I don't have a roommate so it's great.

(S2) Congratulations. Mlao is a great dorm then. It is the largest in chicago. Enjoy!

[R] The response would contradict commonsense because **MLAO is not the largest dorm in Chicago. In fact, there is no dorm in Chicago with that name**. This response makes an unfounded assumption and contradicts common knowledge.

Definition Mismatch

Key Takeaways

- Systematic failures enable targeted mitigation strategies for improvement
 - Context Management: windowed context, turn pairing
 - Instruction Following: in-context examples

Key Takeaways

- Systematic failures enable targeted mitigation strategies for improvement
 - Context Management: windowed context, turn pairing
 - Instruction Following: in-context examples
- GPT is competitive to existing classifiers, surpassing several in performance!
- Promising potential of GPT to perform behavior detection at low-cost

Key Takeaways

- Systematic failures enable targeted mitigation strategies for improvement
 - Context Management: windowed context, turn pairing
 - Instruction Following: in-context examples
- GPT is competitive to existing classifiers, surpassing several in performance!
- Promising potential of GPT to perform behavior detection at low-cost
- Gap between automatic methods and humans for behavior detection

Thank you! Questions?

Paper: <https://github.com/emorynlp/GPT-ABCEval/blob/main/paper.pdf>

GitHub: <https://github.com/emorynlp/GPT-ABCEval>

Contact Info:

sfillwo@emory.edu

<https://www.emorynlp.org>