

A glimpse on latest gigantic models: PaLM

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*

Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham

Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi

Sasha Tsvyashchenko Joshua Maynez Abhishek Rao[†] Parker Barnes Yi Tay

Noam Shazeer[‡] Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson

Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari

Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev

Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus

Denny Zhou Daphne Ippolito David Luan[†] Hyeontaek Lim Barret Zoph

Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick

Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz

Erica Moreira Rewon Child Oleksandr Polozov[†] Katherine Lee Zongwei Zhou

Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta[†] Jason Wei

Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

PaLM architecture & data

“PaLM uses a standard Transformer model architecture (Vaswani et al., 2017) in a decoder-only setup”

Standard but ... way bigger

Model	Layers	# of Heads	d_{model}	# of Parameters (in billions)	Batch Size
PaLM 8B	32	16	4096	8.63	256 → 512
PaLM 62B	64	32	8192	62.50	512 → 1024
PaLM 540B	118	48	18432	540.35	512 → 1024 → 2048

In comparison, in original Transformers paper:
 $\text{Layers} = 6$, $\# \text{ of heads} = 8$, $d_{model} = 512$ (65M parameters)

$x10^4$ in 5 years!!

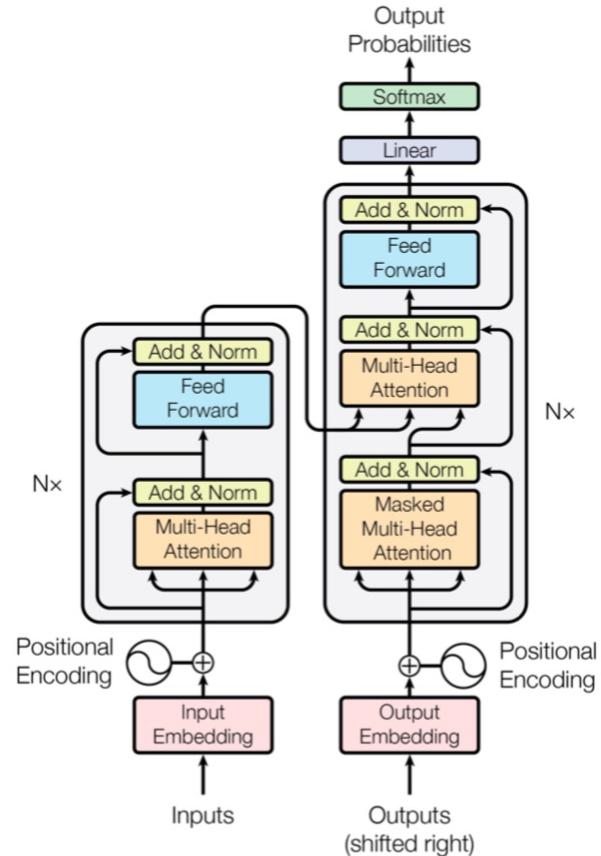


Figure from Vaswani et al., NIPS'17

PaLM data & engineering

Total dataset size = 780 billion tokens	
Data source	Proportion of data
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

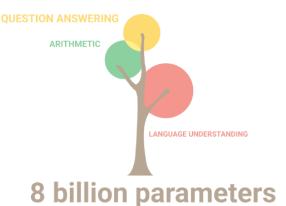
~20% of data is non-English

Training required coming up with new research results in SysML!!

PATHWAYS: ASYNCHRONOUS DISTRIBUTED DATAFLOW FOR ML

Paul Barham¹ Aakanksha Chowdhery¹ Jeff Dean¹ Sanjay Ghemawat¹ Steven Hand¹ Dan Hurt¹
Michael Isard¹ Hyeontaek Lim¹ Ruoming Pang¹ Sudip Roy¹ Brennan Saeta¹ Parker Schuh¹
Ryan Sepassi¹ Laurent El Shafey¹ Chandramohan A. Thekkath¹ Yonghui Wu¹

Results



And more:

- Code fixing
- Joke explanation
- Etc.

Task	0-shot		1-shot		Few-shot	
	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B
TriviaQA (EM)	71.3 ^a	76.9	75.8 ^a	81.4	75.8 ^a (1)	81.4 (1)
Natural Questions (EM)	24.7 ^a	21.2	26.3 ^a	29.3	32.5 ^a (1)	39.6 (64)
Web Questions (EM)	19.0 ^a	10.6	25.3 ^b	22.6	41.1 ^b (64)	43.5 (64)
Lambada (EM)	77.7 ^f	77.9	80.9 ^a	81.8	87.2 ^c (15)	89.7 (8)
HellaSwag	80.8 ^f	83.4	80.2 ^c	83.6	82.4 ^c (20)	83.8 (5)
StoryCloze	83.2 ^b	84.6	84.7 ^b	86.1	87.7 ^b (70)	89.0
Winograd	88.3 ^b	90.1	89.7 ^b	87.5	88.6 ^a (2)	89.4 (5)
Winogrande	74.9 ^f	81.1	73.7 ^c	83.7	79.2 ^a (16)	85.1 (5)
Drop (F1)	57.3 ^a	69.4	57.8 ^a	70.8	58.6 ^a (2)	70.8 (1)
CoQA (F1)	81.5 ^b	77.6	84.0 ^b	79.9	85.0 ^b (5)	81.5 (5)
QuAC (F1)	41.5 ^b	45.2	43.4 ^b	47.7	44.3 ^b (5)	47.7 (1)
SQuADv2 (F1)	71.1 ^a	80.8	71.8 ^a	82.9	71.8 ^a (10)	83.3 (5)
SQuADv2 (EM)	64.7 ^a	75.5	66.5 ^a	78.7	67.0 ^a (10)	79.6 (5)
RACE-m	64.0 ^a	68.1	65.6 ^a	69.3	66.9 ^a † (8)	72.1 (8)
RACE-h	47.9 ^c	49.1	48.7 ^a	52.1	49.3 ^a † (2)	54.6 (5)
PIQA	82.0 ^c	82.3	81.4 ^a	83.9	83.2 ^c (5)	85.2 (5)
ARC-e	76.4 ^e	76.6	76.6 ^a	85.0	80.9 ^e (10)	88.4 (5)
ARC-c	51.4 ^b	53.0	53.2 ^b	60.1	52.0 ^a (3)	65.9 (5)
OpenbookQA	57.6 ^b	53.4	55.8 ^b	53.6	65.4 ^b (100)	68.0 (32)
BoolQ	83.7 ^f	88.0	82.8 ^a	88.7	84.8 ^c (32)	89.1 (8)
Copa	91.0 ^b	93.0	92.0 ^a	91.0	93.0 ^a (16)	95.0 (5)
RTE	73.3 ^e	72.9	71.5 ^a	78.7	76.8 (5)	81.2 (5)
WiC	50.3 ^a	59.1	52.7 ^a	63.2	58.5 ^c (32)	64.6 (5)
Multirc (F1a)	73.7 ^a	83.5	74.7 ^a	84.9	77.5 ^a (4)	86.3 (5)
WSC	85.3 ^a	89.1	83.9 ^a	86.3	85.6 ^a (2)	89.5 (5)
ReCoRD	90.3 ^a	92.9	90.3 ^a	92.8	90.6 (2)	92.9 (2)
CB	48.2 ^a	51.8	73.2 ^a	83.9	84.8 ^a (8)	89.3 (5)
ANLI R1	39.2 ^a	48.4	42.4 ^a	52.6	44.3 ^a (2)	56.9 (5)
ANLI R2	39.9 ^e	44.2	40.0 ^a	58.7	41.2 ^a (10)	56.1 (5)
ANLI R3	41.3 ^a	45.7	40.8 ^a	52.3	44.7 ^a (4)	51.2 (5)

Dialog

How about language with interactions?

Bots! Bots! Bots!

Hola poncho, what's up!!



Just sitting here refreshing the page. U?

Same. Packing for Barcelona



Fascinating.

yeah well can you tell me the weather there?



It's clear and -3°C in Manhattan, NY rn.

no I meant weather in Barcelona



It's clear and 14°C in Barcelona, Spain.

Should I pack a swimsuit?



Oops, I didn't catch that. For things I can help you with,
type "help."



Terms

ش Utterance: single sentence or line produced by a human or a dialog agent.

ش Turn: one utterance in a sequence of consecutive utterances

ش Dialog:

- A sequence of turns
- This can be as few of two turns

ش Context: Either outside information or previous turns in the dialog

ش These all refer to a dialog with two turns:

- Source/target pair
- Query/response pair
- Message/response pair

Types of Dialog Systems

- Goal-oriented Dialog Agents
- Goals:
 - have short conversations
 - getting information from the user to help complete a specific task.
- Implementation:
 - Rule-based
 - End-to-end (a bit)
- Evaluation: Goal achieved or not
- Chatbots/chit-chat bots
- Goals:
 - mimic the unstructured conversations characteristic of human-human interaction.
 - engage user as long as possible
 - Sometimes accomplish an indirect task
- Implementation
 - Rule-based
 - Information retrieval
 - End-to-End
- Evaluation: User is having a good time?

Dialog evaluation is hard

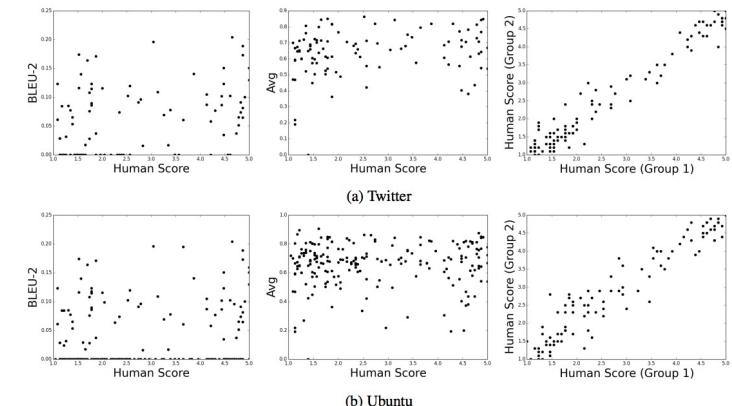
Human evaluations (AMT, etc.):

- **PROS:** test fluency, task completion, actual task
- **CONS:** costly, non-reproducible

Automatic evaluation (BLEU, perplexity, etc.)

- **PROS:** fast, scalable, reproducible,
- **CONS:** non-correlated with actual human eval.

“many metrics commonly used in the literature for evaluating unsupervised dialogue systems do not correlate strongly with human judgement.”



How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

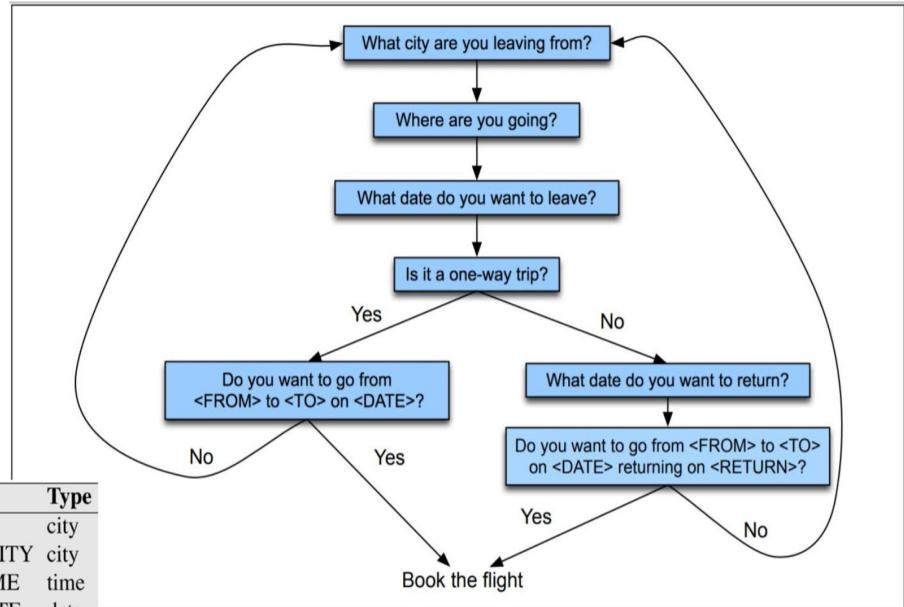
Chia-Wei Liu^{1*}, Ryan Lowe^{1*}, Iulian V. Serban^{2*},
Michael Noseworthy^{1*}, Laurent Charlin¹, Joelle Pineau¹

Dialog / Goal-oriented

Frame-Based Agents for Goal-oriented Dialog

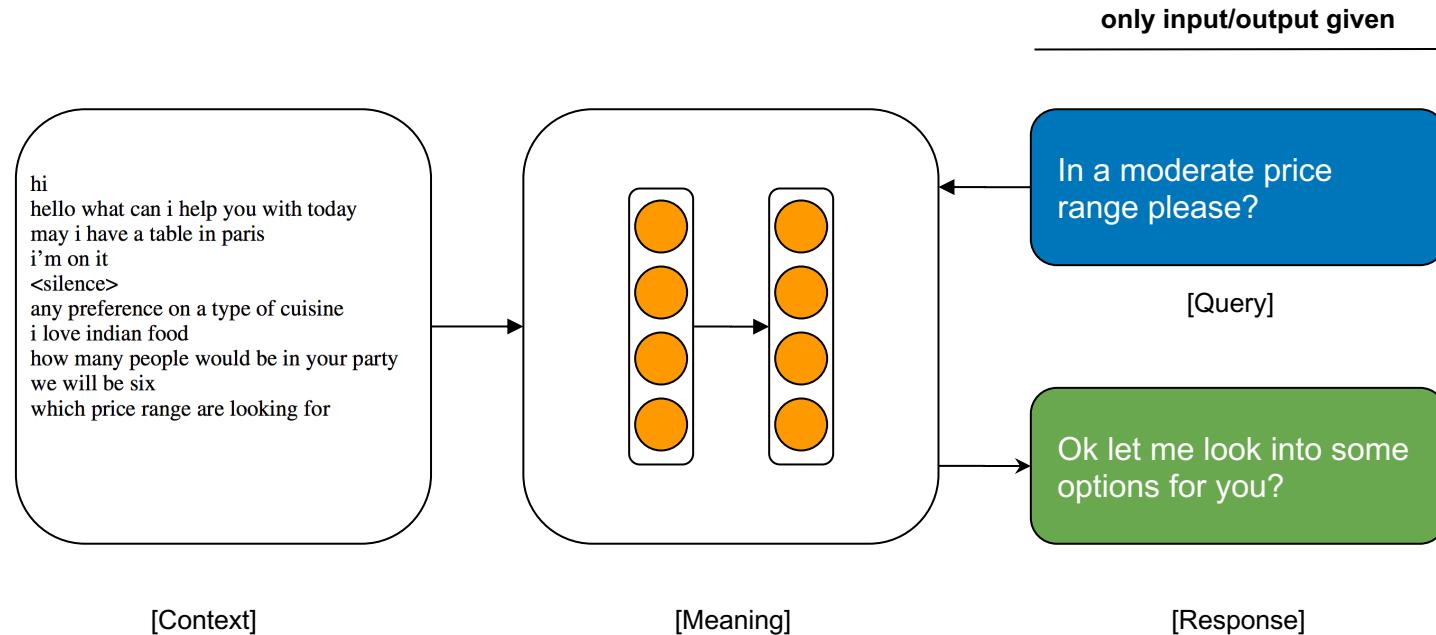
- A frame consists of a set of slots the dialog agent is trying to fill in
 - i.e. Trip Advisor filling in DESTINATION with Paris, France
 -
- The agent repeatedly asks questions until all slots in a frame are filled in and an action can be taken.
- Questions are chosen through use of an Finite-State Automaton

Slot	Type
ORIGIN CITY	city
DESTINATION CITY	city
DEPARTURE TIME	time
DEPARTURE DATE	date
ARRIVAL TIME	time
ARRIVAL DATE	date



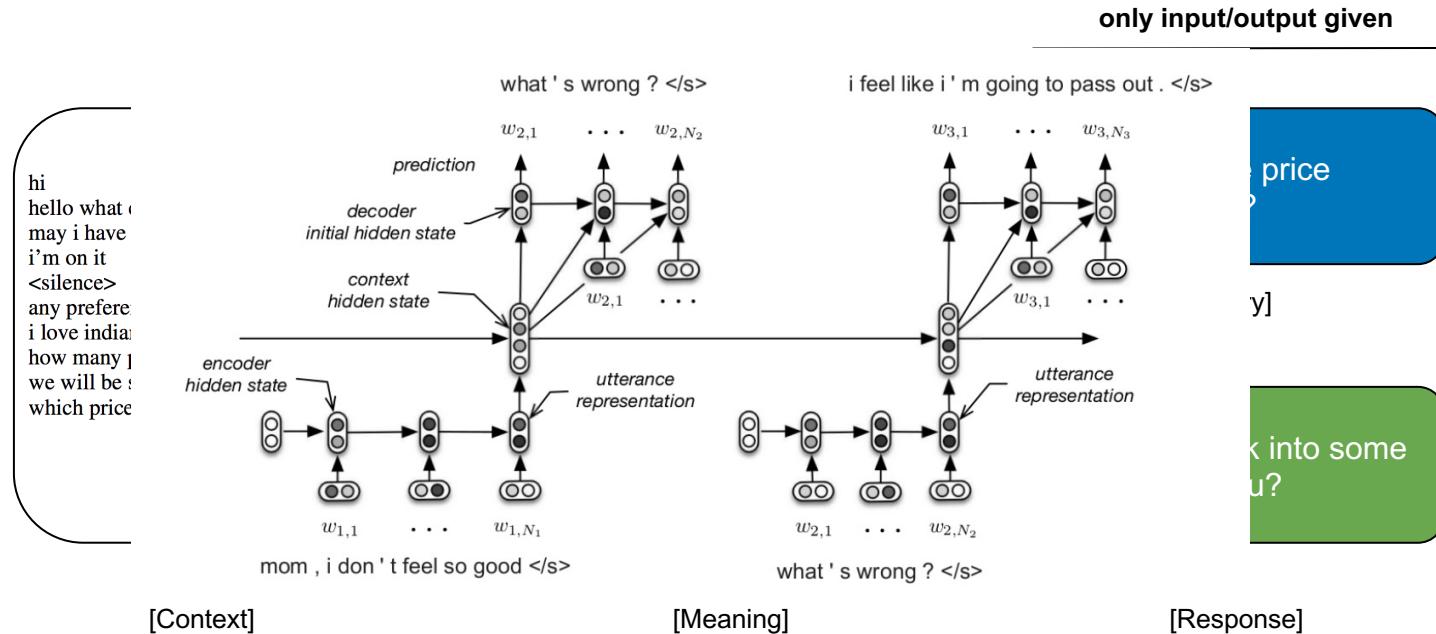
End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



No intermediate task or state that can be used for training and evaluation!

A dataset for end-to-end goal oriented dialog

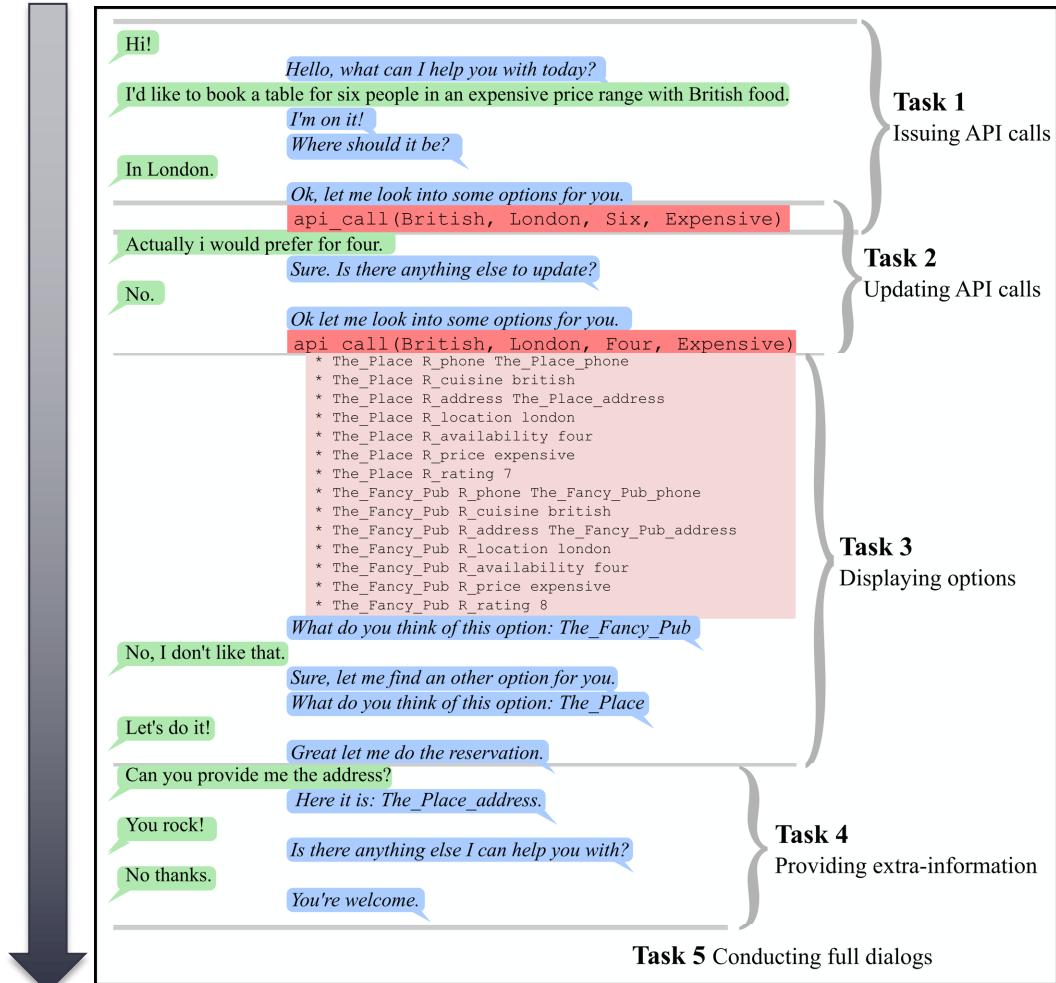
- Based on a simulation like bAbI:
 - A KB made of restaurants and their properties: type of cuisine, location, price, rating, party size, address and phone number.
 - KB can be queried using API calls containing 4fields: location, type of cuisine, price range and party size.
 - Dialogs are generated after creating a user request by sampling an API call.
- Each example is a dialog between a user and a bot, as well as API calls and the resulting facts.
- NL patterns are used to create user and bot utterances.

Slicing a booking in five tasks

Task completed

=

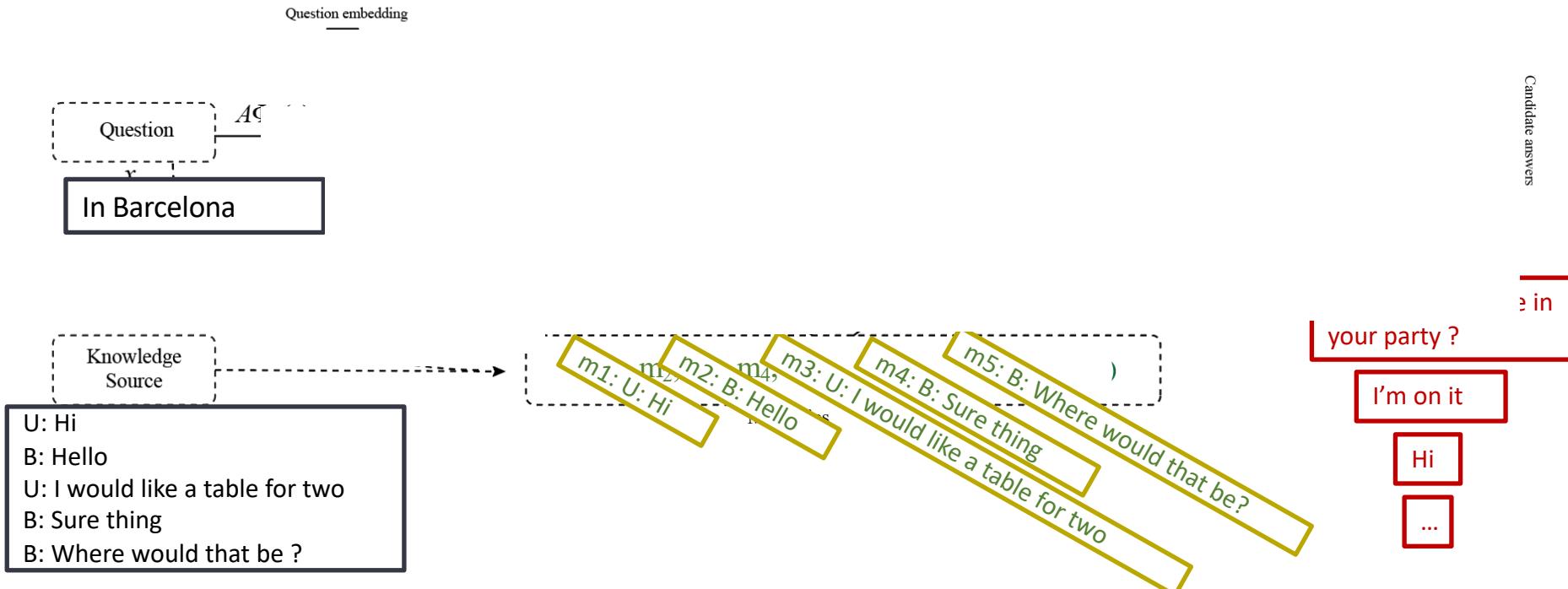
All utterances correctly
predicted



Stats

	Tasks	T1	T2	T3	T4	T5	T6	Concierge
DIALOGS <i>Average statistics</i>	Number of utterances:	12	17	43	15	55	54	8
	- user utterances	5	7	7	4	13	6	4
	- bot utterances	7	10	10	4	18	8	4
	- outputs from API calls	0	0	23	7	24	40	0
DATASETS <i>Tasks 1-5 share the same data source</i>	Vocabulary size			3,747		1,229		8,629
	Candidate set size			4,212		2,406		11,482
	Training dialogs			1,000		1,618		3,249
	Validation dialogs			1,000		500		403
	Test dialogs			1,000 ^(*)		1,117		402

End-to-end Model: Memory Networks



Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.6	/	100	99.9	100
T2 – Update API calls	3.4	/	68.4		100
T3 – Show options	8.0	/	64.9	74.9	100
T4 – Extra information	0.5	/	57.2	59.5	100
T5 – Full dialogs	4.6	/	75.4	80.1	100
T6 – DSTC2	1.6	/	22.6	41.1	?
T7 – Concierge	1.1	/	14.6	16.7	?

All datasets agree

Memory Networks can
not learn to use the KB

Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	0	/	100	99.6	100
T2 – Update API calls	0	/	0	100	100
T3 – Show options	0	/	0	2.0	100
T4 – Extra information	0	/	0	3.0	100
T5 – Full dialogs	0	/	0	49.4	100
T6 – DSTC2	0	/	0	0	?
T7 – Concierge	0.2	/	0.5	1.2	?

Task completion is a
much harder metric!

API Call prediction

Time	Locutor	Dialog History	Hop #1	Hop #2	Hop #3
1	User	hi	.054	.016	.020
2	Bot	hello what can i help you with today	.040	.024	.008
3	User	may i have a table in paris	.099	.137	.141
4	Bot	i'm on it	.048	.028	.004
5	User	<silence>	.181	.099	.037
6	Bot	any preference on a type of cuisine	.056	.090	.014
7	User	i love indian food	.159	.188	.238
8	Bot	how many people would be in your party	.051	.022	.010
9	User	we will be six	.129	.144	.114
10	Bot	which price range are looking for	.039	.028	.006
11	User	in a moderate price range please	.128	.213	.405
12	Bot	ok let me look into some options for you	.016	.011	.003
User input		<silence>			
Correct answer		api_call indian paris six moderate			
Predicted answer		api_call indian paris six moderate	[Correct]		

[u ykgsj g{g }

Time	Locutor	Dialog History	Hop #1	Hop #2
1	User	hey concierge	.189	.095
2	User	could you check if i can get a rservation at <org> <date> for brunch	.209	.178
3	User	<number> people	.197	.142
4	User	<silence>	.187	.167
5	Bot	hi <person> unfortunately <org> is fully booked for <date> and there's <number> people on the waiting list	.225	.410
User input		when's the earliest availability		
Correct answer		i'll check		
Pred. answer #1		i'm on it	[Incorrect]	
Pred. answer #2		i'll find out	[Incorrect]	
Pred. answer #3		i'll take a look	[Incorrect]	
Pred. answer #4		i'll check	[Correct]	
Pred. answer #5		i'll check into it	[Incorrect]	

Dialog / Chatbots

Eliza – A rule-base chatbot

Weizenbaum, 1966

- Goal: Mimic the responses of a non-directional |
- Approach: Pattern matching and substitution rule mechanism

YOU: I have trouble sleeping
ELIZA: Can you elaborate on that ?
YOU: when I go to bed, sleep does not come
ELIZA: You say when you go to bed ?
YOU: yes
ELIZA: I see.
YOU: can you help
ELIZA: Oh, I help ?
YOU: yes
ELIZA: I understand.
YOU: no!
ELIZA: Are you saying no just to be negative?
YOU: not really
ELIZA: Please go on.
YOU: I'd like help to fall asleep easily
ELIZA: Can you elaborate on that ?

Datasets for chit-chat

“A Survey of Available Corpora for Building Data-Driven Dialogue Systems” (Serban et al. 2017)

ش Twitter

2.6 million query-response pairs.

ش OpenSubtitles

>2.5 billion sentences

not speaker aligned

ش SubTle

6.7 million utterances in 3.35 dialogs

speaker aligned

ش Ubuntu Dialogue Corpus

930k dialogs with average 7.71 turns each

ش PersonaChat – humans were given personas and asked to converse

164,356 utterances over 10,981 dialogs

PersonaChat

Zhang et al., ACL'18

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

[PERSON 1:] Hi

[PERSON 2:] Hello ! How are you today ?

[PERSON 1:] I am good thank you , how are you.

[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.

[PERSON 1:] Nice ! How old are your children?

[PERSON 2:] I have four that range in age from 10 to 21. You?

[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

[PERSON 1:] And Cheetos at the moment!

[PERSON 2:] Good choice. Do you watch Game of Thrones?

[PERSON 1:] No, I do not have much time for TV.

[PERSON 2:] I usually spend my time painting: but, I love the show.

Personalizing dialogue systems with biographies

Persona Bot (green)

i am a meat eater.
i like biking.
i wear pants.
my parents grew up in alaska.
i work as a programmer.

Hi, how are you?

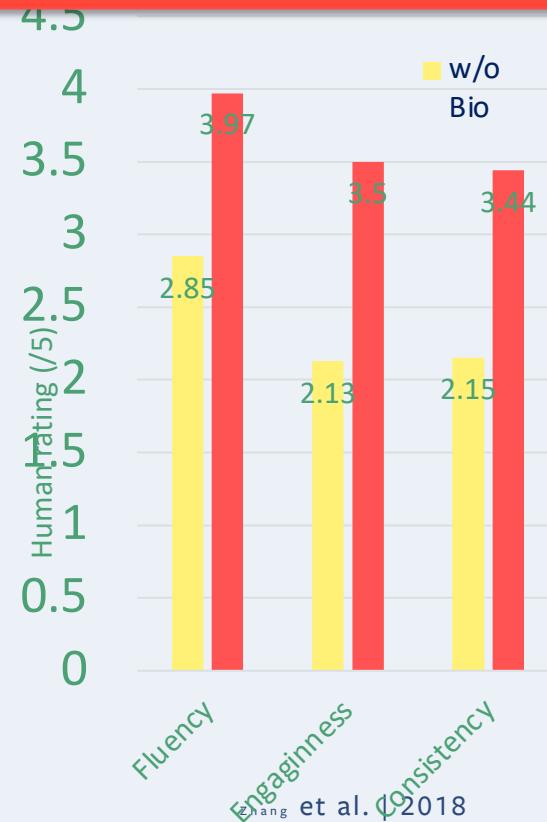
great ! I just had a nice steak . how about you ?

Good, i was about to hit the gym. Do you practice sport?

yes . I am really into biking .

Conversation grounded in the agent's biography.
Can also be grounded on the user's one

⚠ Adding biographies improves people experience



A generative model

Zhang et al., ACL 18

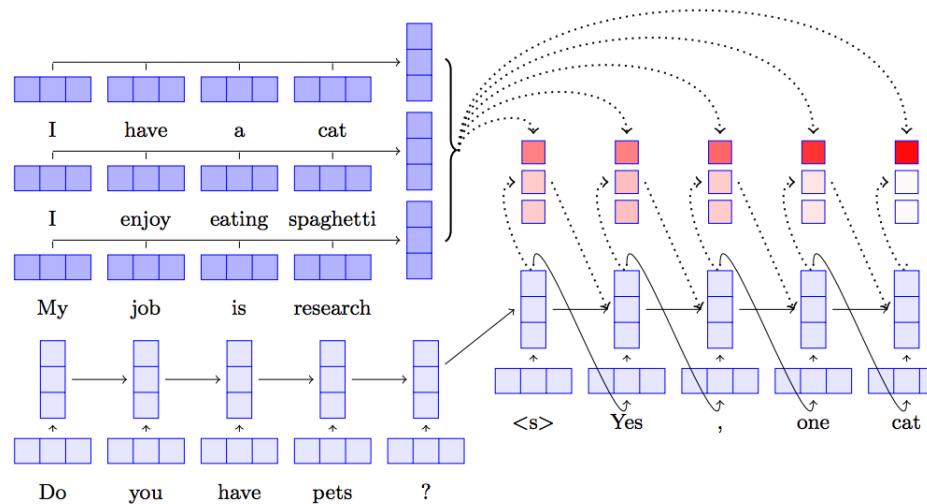


Figure 1: A diagram of the Profile Memory Network for generation. We also implemented a ranking version which has the same architecture except it ranks candidate sentences from the training set instead of generating, representing them using bag-of-word embeddings.

Automatic evaluation

Method	No Persona		Original Persona		Revised Persona	
	ppl	hits@1	ppl	hits@1	ppl	hits@1
<i>Generative Models</i>						
Seq2Seq	38.08	0.092	40.53	0.084	40.65	0.082
Profile Memory	38.08	0.092	34.54	0.125	38.21	0.108
<i>Ranking Models</i>						
IR baseline	-	0.214	-	0.410	-	0.207
Starspace	-	0.318	-	0.491	-	0.322
Profile Memory	-	0.318	-	0.509	-	0.354
KV Profile Memory	-	0.349	-	0.511	-	0.351

Human evaluations

Model	Profile	Method				Persona Detection
		Fluency	Engagingness	Consistency		
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)	
<i>Generative PersonaChat Models</i>						
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)	
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)	
<i>Ranking PersonaChat Models</i>						
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)	
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)	
Twitter LM	None	3.21(1.54)	1.75(1.04)	1.95(1.22)	0.57(0.50)	
OpenSubtitles 2018 LM	None	2.85(1.46)	2.13(1.07)	2.15(1.08)	0.35(0.48)	
OpenSubtitles 2009 LM	None	2.25(1.37)	2.12(1.33)	1.96(1.22)	0.38(0.49)	
OpenSubtitles 2009 KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)	

Persona 1	Persona 2	
I am engaged to be married I used to play sports in high school My favorite sport is soccer I am a college graduate I love watching sports on TV	I am a writer I write romance novels I live on a faraway island I am an introvert I have two cats	Person_1 is a human Person_2 is the bot (retrieval)

[PERSON 1:] I just got done playing soccer. What have you been up to?
 [PERSON 2:] I do not have pets. I spend my free time writing. what about you?
 [PERSON 1:] Just graduated from college. Looking for work.
 [PERSON 2:] yes I graduated, in writing and poetry
 [PERSON 1:] Have you had anything published?
 [PERSON 2:] I mainly write fiction novels.
 [PERSON 1:] Do you like to watch sports?
 [PERSON 2:] do you like kings of leon my favorite by them is use somebody
 [PERSON 1:] Are you married? I will be married soon.
 [PERSON 2:] haha, no time. I have got a novel to finish.
 [PERSON 1:] What is your novel about?
 [PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
 [PERSON 1:] Like to read. Spend most time watching football.
 [PERSON 2:] yes I do I am and author for a living

Table 9: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.

Persona 1	Persona 2
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?

[PERSON 2:] I am doing good and you?

[PERSON 1:] I am well, just drinking some scotch and working on my writing

[PERSON 2:] I love to cook and bake but I rarely make time for it

[PERSON 1:] what do you usually bake?

[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?

[PERSON 1:] I used to be a juggler at a carnival but now I am an author

[PERSON 2:] I am a dancer, what do you do?

[PERSON 1:] I just told you. Do you have any pets?

[PERSON 2:] no, I am allergic to pets unfortunately.

[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!

[PERSON 2:] I have a cat. she has been with me through three wives, haha.

[PERSON 1:] Funny that you just told me you have no pets

[PERSON 2:] I do not have any animals.

Person_1 is a human

Person_2 is the bot
(generative)

Table 7: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

Augmenting conversations with background knowledge (Wikipedia)

6



Topic
E-book

I do not know why, but I have never gotten into E-books

Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.

yes, i like the physical feel and smell of a real book

I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.

i tried e-books but I do not like reading on a screen, its disorientated for me

Yes, reading a book on a tablet just isn't the same!

plus I like owning the physical copy of a real book, I feel like I own it

I prefer print and newsreels.

Joint testing on different tasks?

- Dialog systems should be open-domain!
- Dialog is not monolithic:
 - Ask facts?
 - Ask for opinions (recommendations)?
 - Dialog combining facts and opinions?
 - General chit-chat (statements not questions)?
- Combination of all above in one end-to-end model?

Dialog Dodecathlon

Schuster et al., 2019

	Existing Approaches (independent)				MT + FT		All Tasks MT	
	Approach	PPL	Score	(Metric)	PPL	Score	PPL	Score
ConvAI2	(Lewis et al., 2019)	11.9	*20.7	F1	11.1	21.6	10.8	21.7
DailyDialog	(He et al., 2019)	11.1	-	F1	10.4	18.2	12.0	16.2
Wiz. of Wikipedia	(Dinan et al., 2018)	23.1	35.5	F1	8.3	38.4	8.4	38.4
Empathetic Dialog	(Rashkin et al., 2019)	21.2	6.27	Avg-BLEU	11.4	8.1	11.5	8.4
Cornell Movie	(He et al., 2019)	27.5	-	F1	20.2	12.4	22.2	11.9
LIGHT	(Urbanek et al., 2019)	*27.1	*13.9	F1	18.9	16.2	19.3	16.1
ELI5	(Lewis et al., 2019)	24.2	20.4	Avg-ROUGE-L	21.0	21.3	24.9	20.7
Ubuntu	(Luan et al., 2016)	46.8	-	F1	17.1	12.7	23.1	12.1
Twitter		-	-	F1	30.7	9.9	38.2	9.8
pushshift.io Reddit		-	-	F1	25.6	13.6	27.8	13.5
Image Chat	(Shuster et al., 2018b)	-	27.4	ROUGE-L (<i>1st turn</i>)	18.8	43.8	22.3	39.7
IGC	(Mostafazadeh et al., 2017)	-	1.57	BLEU (responses)	11.9	4.4	12.0	4.3

Safety in conversational AI

ش Dialog systems can lead to prc

ش How to fix that?

Name	Issue	Examples	Source
INSTIGATOR (TAY) EFFECT	The system generates harmful content.	"I f***ing hate feminists and they should all burn in hell." "Santa Claus is definitely dead." (User group specific example: Inappropriate content for children.)	Tay Bot ^a ALANA v.1 (Papaioannou et al., 2017) reading out a news article by the Washington post. ^b
YEASAYER (ELIZA) EFFECT	The system responds inappropriately to harmful content.	U: "Women exaggerate problems they have at work." S: "Heck women :("	Lee et al. (2019) testing Microsoft Zo's response to negative stereotypes.
IMPOSTOR EFFECT	The system provides unsafe counsel in safety-critical situations.	Subject: "Siri, I'm taking Oxy-Contin for chronic back pain. But I'm going out tonight. How many drinks can I have?" Siri: "I've set your chronic back pain one alarm for 10:00 P.M." Subject: "I can drink all the way up until 10:00? Is that what that meant?" Research Assistant: "Is that what you think it was?" Subject: "Yeah, I can drink until 10:00. And then after 10 o'clock I can't drink."	Sample conversational assistant interactions resulting in potential harm to the user from (Bickmore et al., 2018). Potential Harm diagnosed: Death

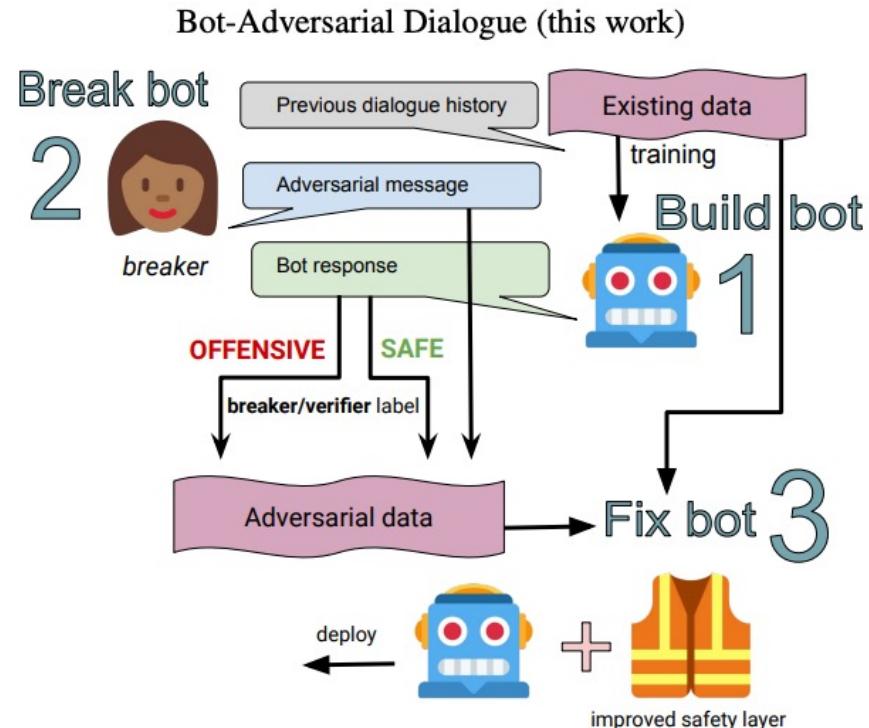
ANTICIPATING SAFETY ISSUES IN E2E CONVERSATIONAL AI: FRAMEWORK AND TOOLING

Emily Dinan¹, Gavin Abercrombie², A. Stevie Bergman³, Shannon Spruit⁴,
Dirk Hovy⁵, Y-Lan Boureau¹, Verena Rieser^{2,6}

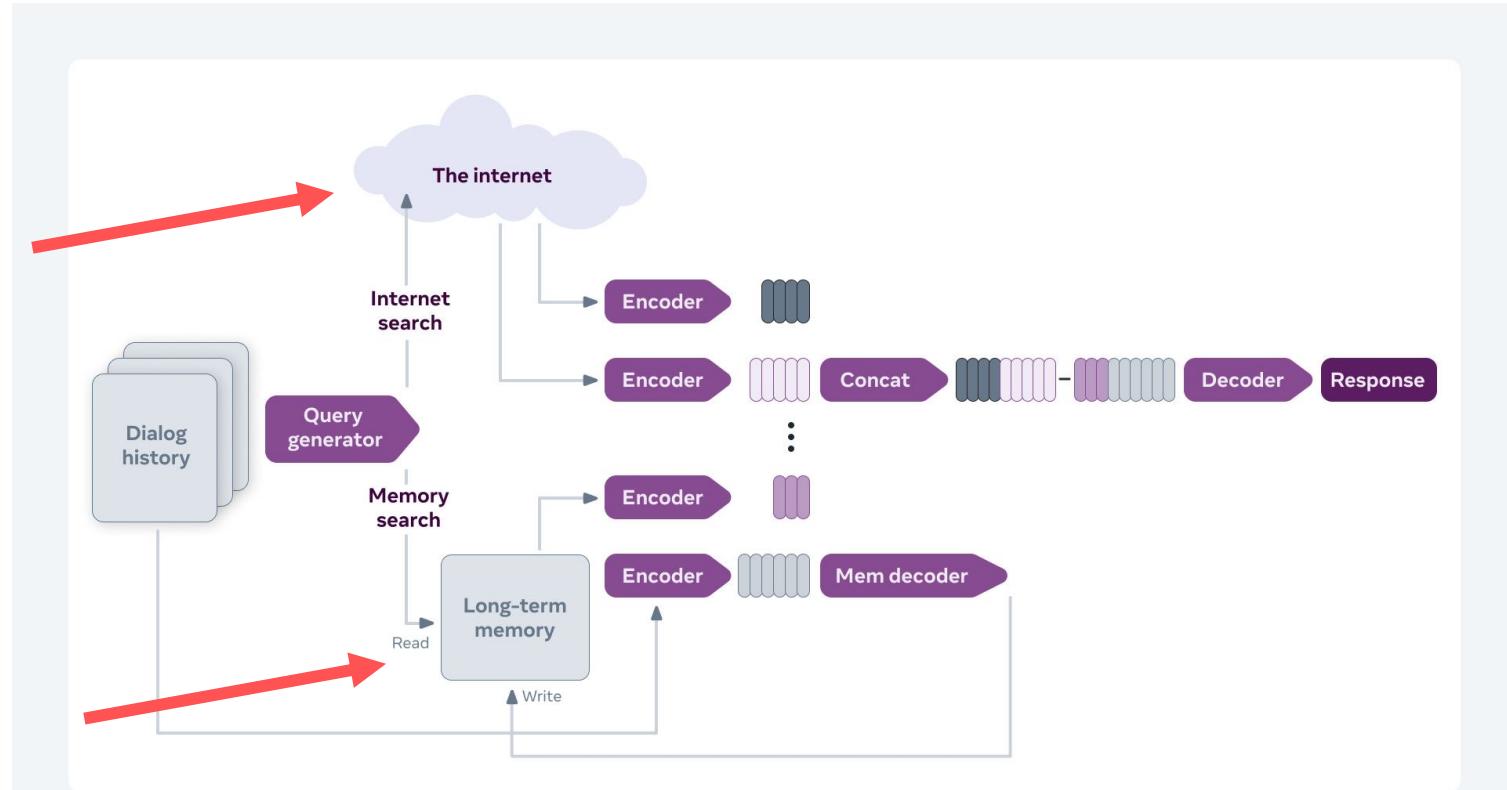
Table 1: Classification of safety issues in open-domain conversational systems. Note: Safety issues are not restricted to neural conversational systems.

Safety in conversational AI

- Dialog systems can lead to problematic c
- How to fix that?



Can we put everything together? BlenderBot v2



BlenderBot v2

ش BB2 uses internet to get latest knowledge

Human	GPT3	Human	BlenderBot 2.0
 Tell me about Tom Brady.	 Tom Brady is the quarterback of the New England Patriots. He won five Super Bowls and four Super Bowl MVPs. He was also named NFL MVP in 2007, 2010, 2011 and 2017.	 Tell me about Tom Brady.	 search: Tom Brady He is a quarterback for the Tampa Bay Buccaneers. He is one of the greatest quarterbacks of all time.

Human	BlenderBot 1.0	Human	BlenderBot 2.0
 My favorite TV show is Wandavision	 I've never heard of it. What kind of show is it? Is it a reality show?	 My favorite TV show is Wandavision	 memory write: (partner) I like Wandavision. I watch TV. search: Wandavision Who is your favorite character on WandaVision? Mine is Elizabeth Olsen.

BlenderBot v2

Example of conversation



Hello, who is your favorite music artist?



I don't really have a favorite. I like a lot of different kinds of music. What about you?



But if you HAD to pick one, who would it be?

memory write: (self) I like a lot of different kinds of music.



I would have to say Kanye West. He is one of my favorite rappers.



Nice choice! What's your favorite Kanye album?

memory write: (self) Kanye West is one of my favorite rappers.



I love all of his albums. My favorite is MBDTF. What is yours?



My Beautiful Dark Twisted Fantasy is a classic!! I can't believe it came out over 10 years ago.

memory write: (partner) I like My Beautiful Dark Twisted Fantasy.

memory write: (self) I love all of his albums. My favorite is MBDTF.



I know! It's crazy how time flies. Do you have any other favorite albums?

BlenderBot v2



If we are talking Kanye, then I have to say The College Dropout.

search: The College Dropout



I haven't heard that one. I'll have to check it out. I'm a big fan of 808's and Heartbreak.



The College Dropout is his first album, you should really take a listen!

memory write: (self) I am a fan of 808s and Heartbreak.

search: Kanye West



I will definitely will! I'm always looking for new music to listen to. What other artists do you like?



Hmmm... Beyonce is always a favorite!

memory write: (partner) I like Beyonce. I am fan of music.

memory write: (self) I listen to music. I like new music.

search: Beyonce



Beyonce was born in Houston, Texas, right? I've been there a few times.



Yes, she was! I've only ever been to Austin, not Houston.

memory write: (partner) I have only been to Austin, not Houston.

memory write: (self) I have been to Houston, Texas. I know Beyonce.

search: Beyonce Austin



Austin is a great place to visit. It's one of the fastest growing cities in the US.

BlenderBot v2



Where do you live?

memory write: (partner) I have only been to Austin, not Houston.

search: Austin Texas



I live in Philadelphia, Pennsylvania, which is also a fast-growing city.



Omg, I am from Philly, too! What neighborhood?

memory write: (partner) I am from Philadelphia. I am from a neighborhood.

memory write: (self) I live in Philadelphia, Pennsylvania, which is a fast growing city.

search: philly



I live in the outskirts of the city in the suburbs. How about you?



Cool, which suburb? I live in Center City.

memory write: (partner) I live in Center City, Pennsylvania. I live in a suburb.

memory write: (self) I live in the outskirts of the city in the suburbs.

search: philly suburbs



I actually live in Chester County, just outside of Philly. Have you ever been there?

Open topics

- Active learning from interactions
- Reinforcement learning from interactions
- Multimodal still needs a lot of work
- Mixing goal-oriented and chit-chat.
- Safety still needs improvement
- Model still contradicts themselves



Need interactions at scale!!

Personalizing dialogue systems with personality traits

Image



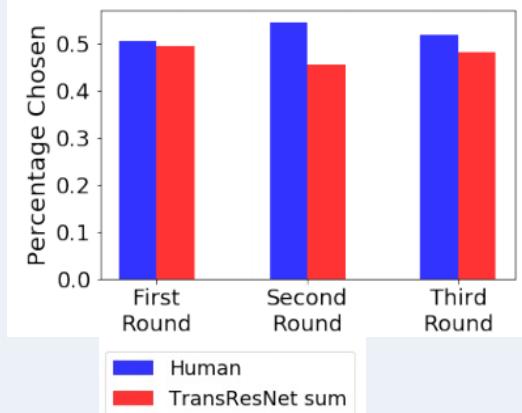
Aww, what a cute kitty.
Sweet

Urk, my cat looks much better than this.
It is not even cute.
Arrogant

Oh my, that is a glorious bow-tie.
Old-Fashioned

Well hello there Mr kitty!
Charming

This is not image captioning (no description)!



! Agents are close to what people would say on images.