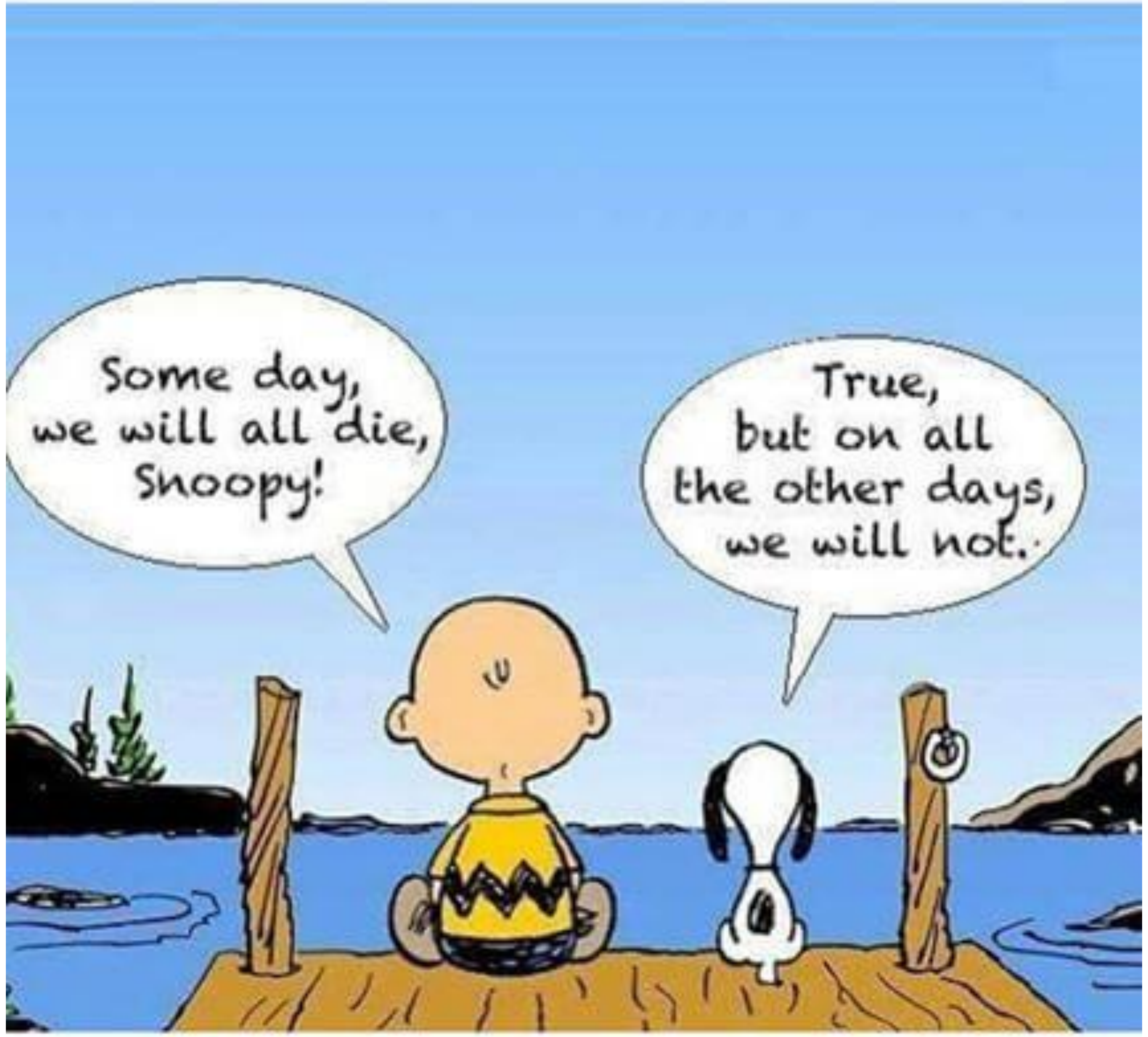
Learning End-to-End Goal-Directed Dialog

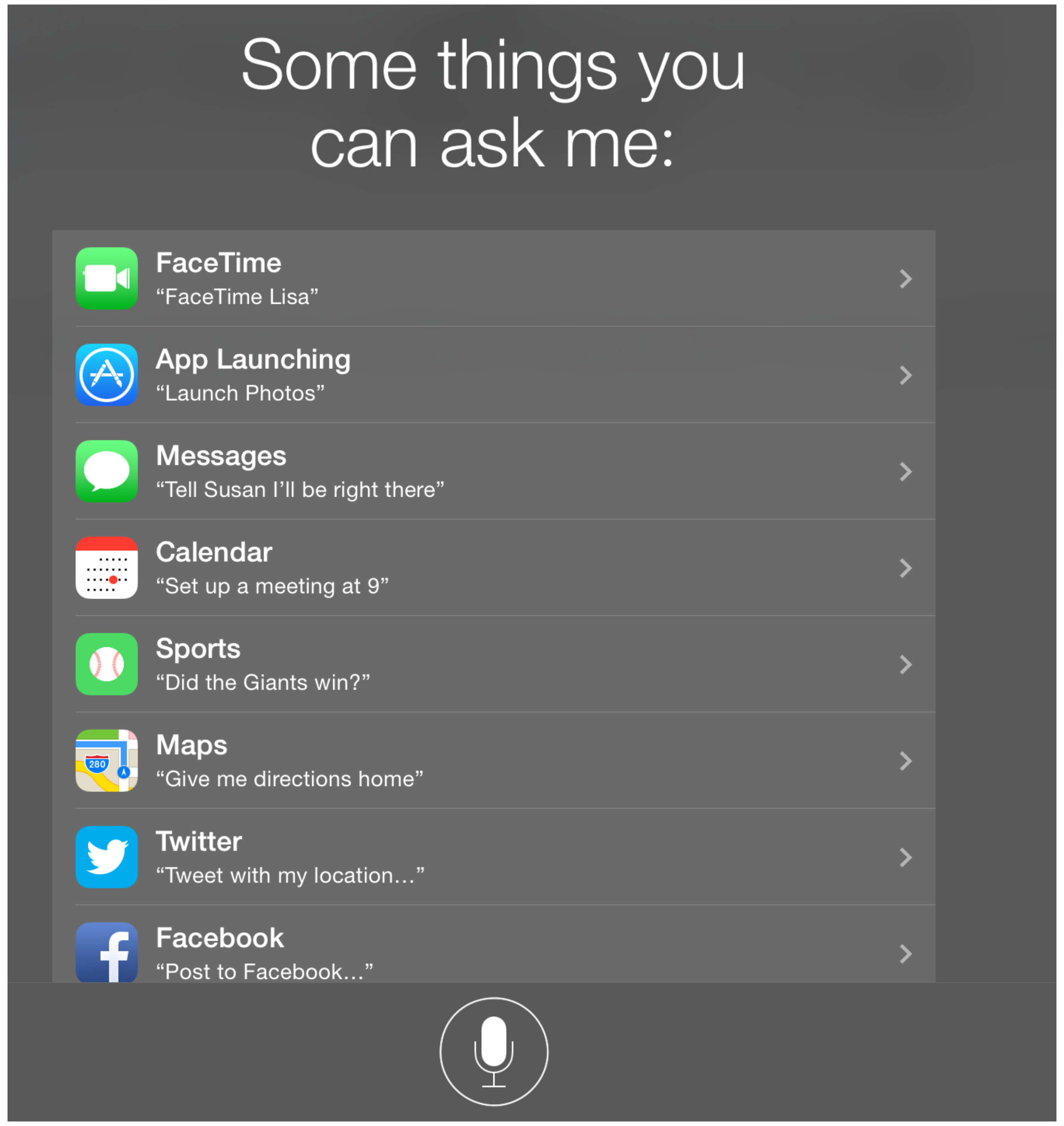
Antoine Bordes, Y-Lan Boureau, Jason Weston

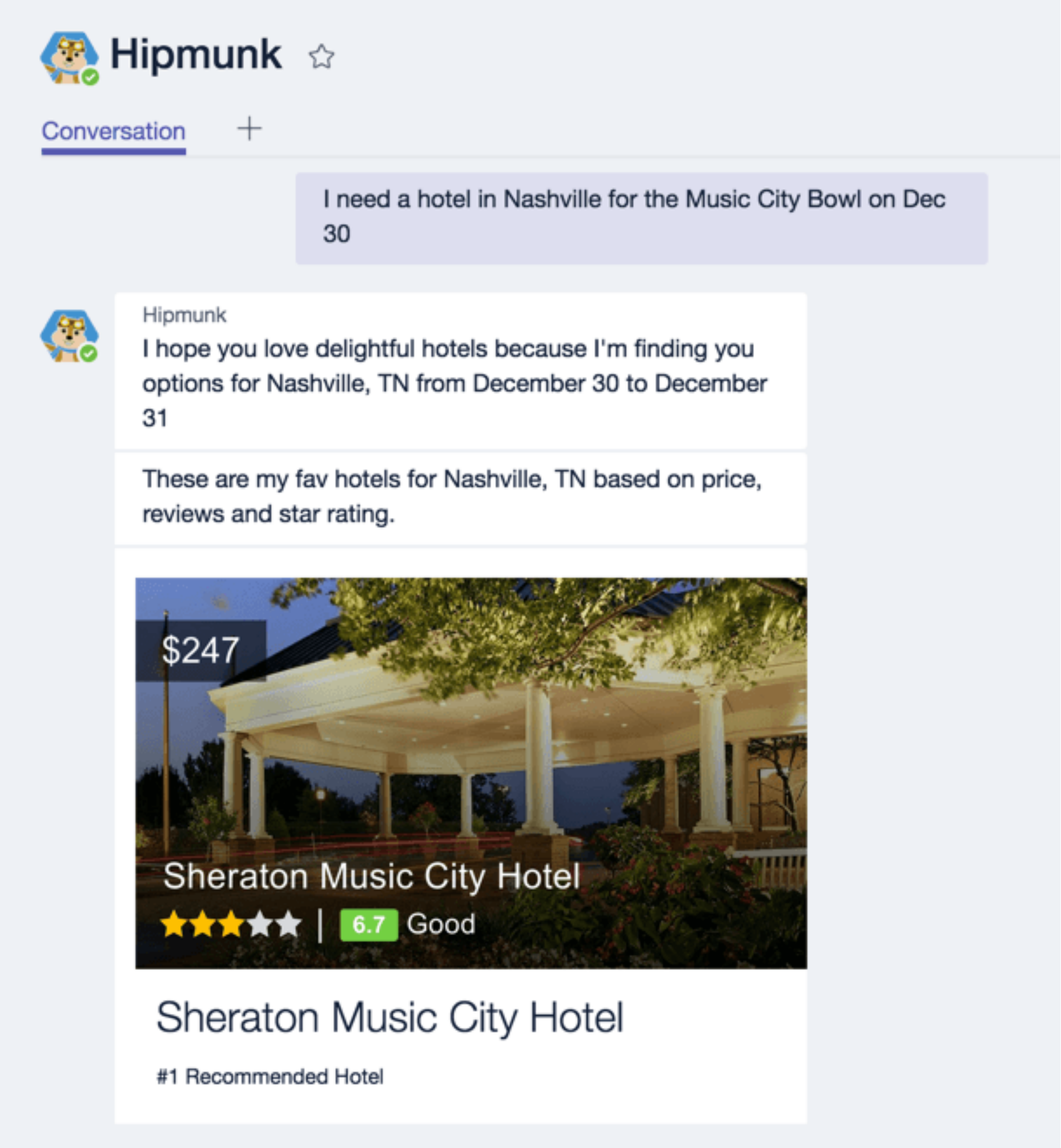
Facebook AI Research

Dialog: aspirations and practice



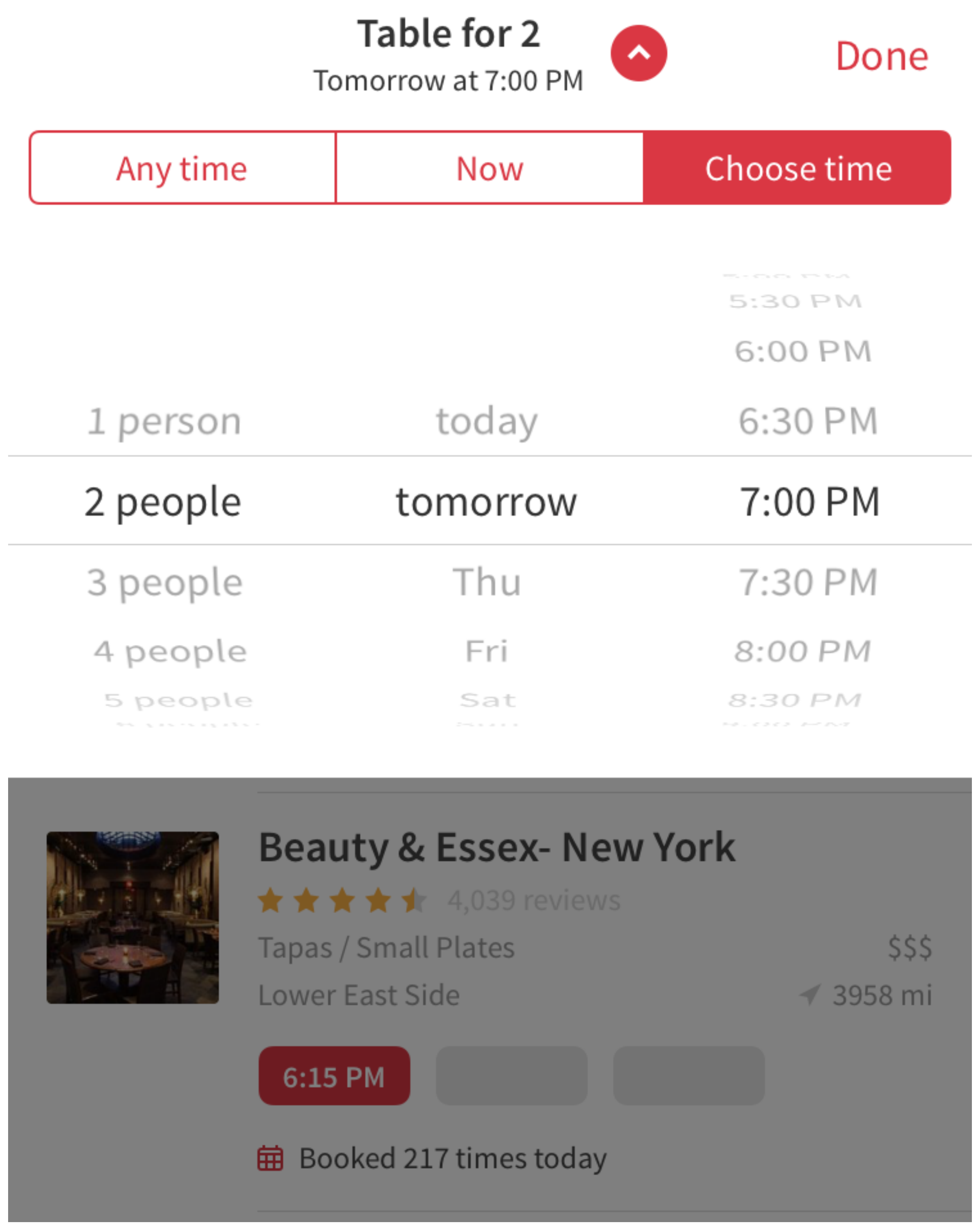
Dialog: aspirations and practice





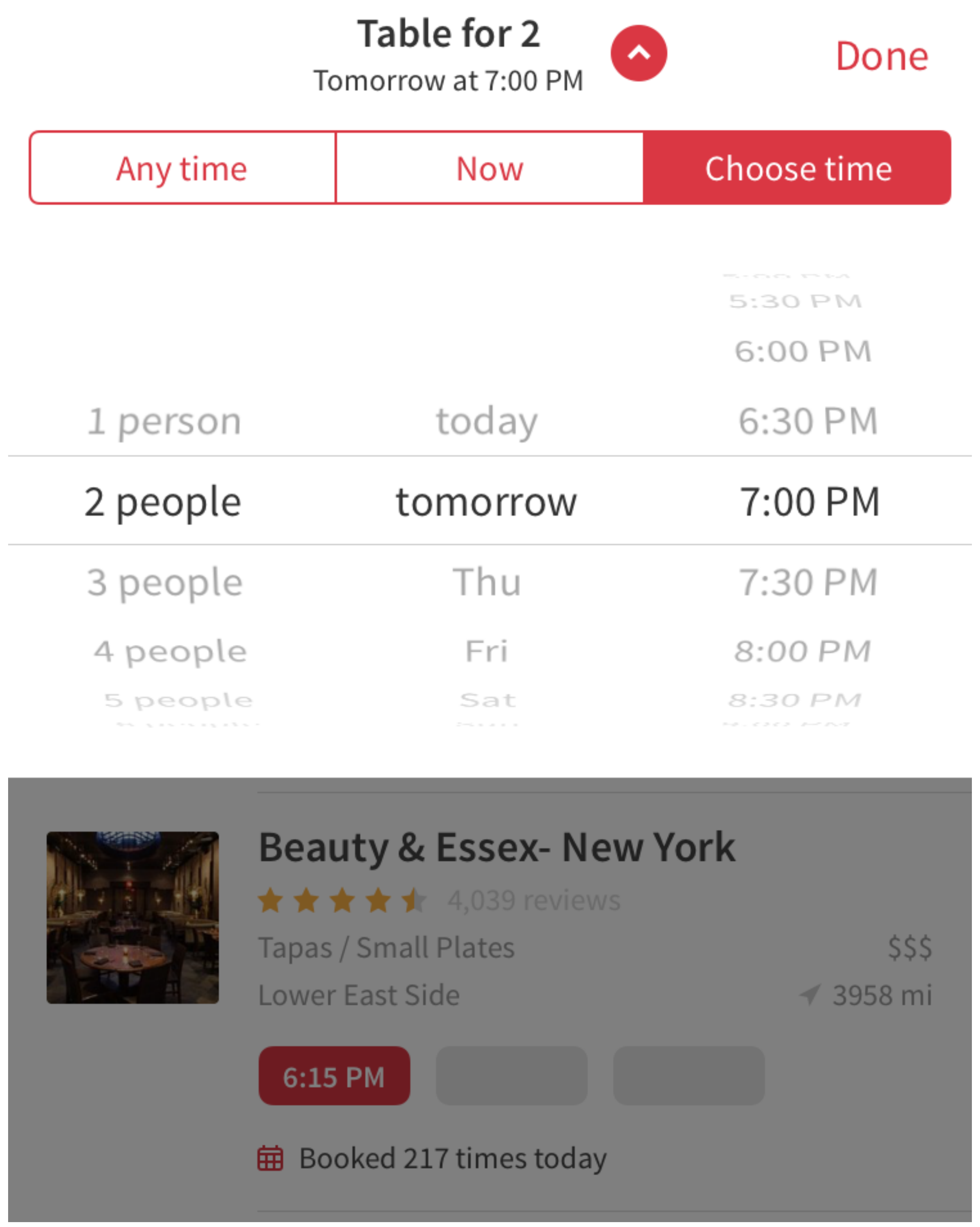
Traditional dialog systems

Slot filling *(Lemon et al., 2006; Wang and Lemon, 2013; Young et al., 2013)*

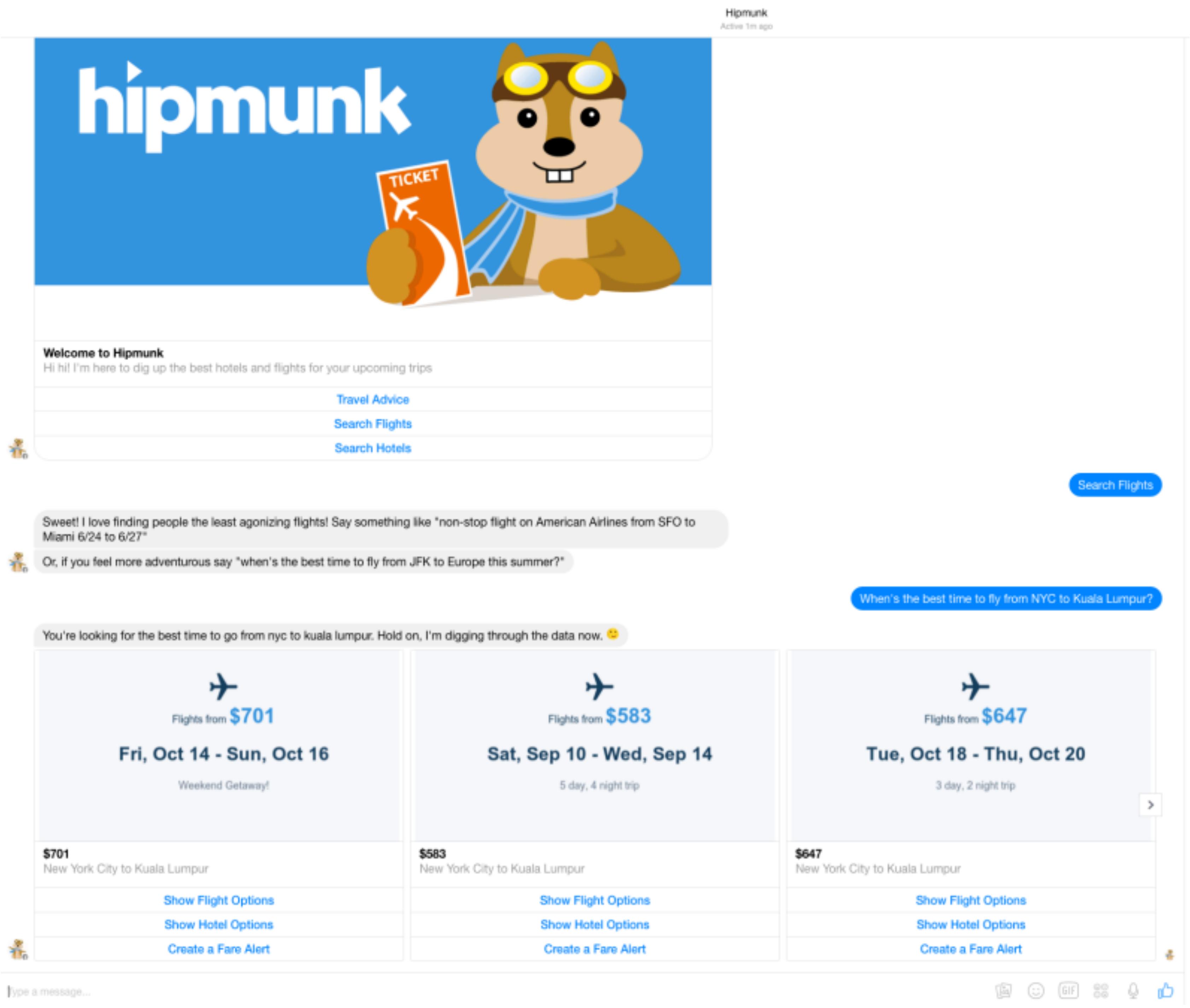


Traditional dialog systems

Slot filling *(Lemon et al., 2006; Wang and Lemon, 2013; Young et al., 2013)*



???



End-to-End dialog

- Usually based on neural networks (Shang *et al.*, 2015; Vinyals and Le, 2015; Sordoni *et al.*, 2015; Serban *et al.*, 2015a; Dodge *et al.*, 2016)
- No assumption on domain or dialog state structure:
  - Easy to automatically scale up to new domains
- Promising performance in non-goal-oriented chit-chat settings
  - Trained to predict next occurrence in social media and forum threads (Ritter et al 2011, Wang et al 2013, Lowe et al 2015) or movie conversations (Banchs, 2012)

An open resource to test end-to-end goal-oriented dialog

- Easy diagnostic analysis: breaking down objective into subtasks
- Easy reproducibility and comparisons
- Lightweight and easy to use

https://fb.ai/babi

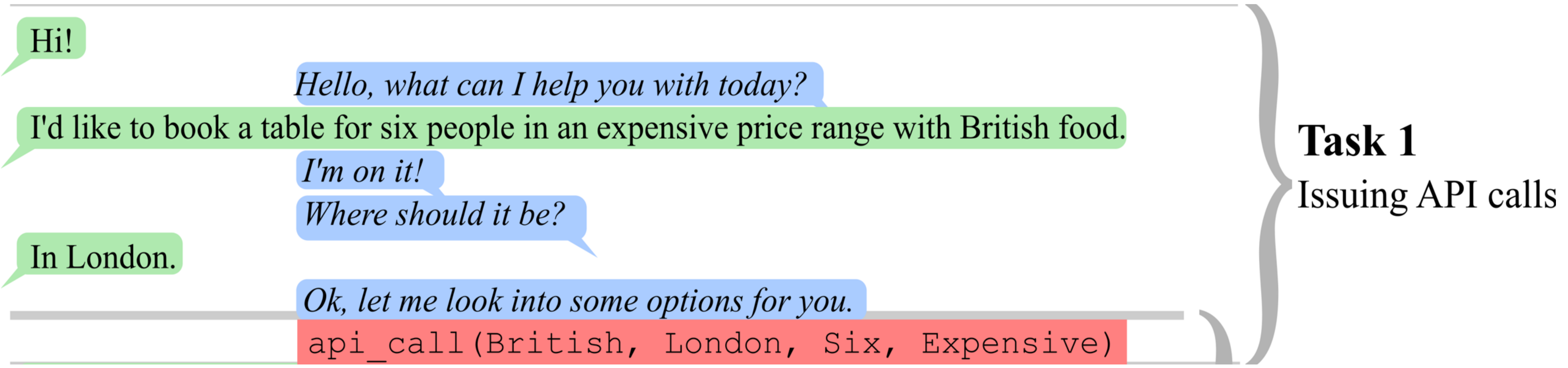
An open resource to test end-to-end goal-oriented dialog

- Easy diagnostic analysis: breaking down objective into subtasks
- Easy reproducibility and comparisons
- Lightweight and easy to use

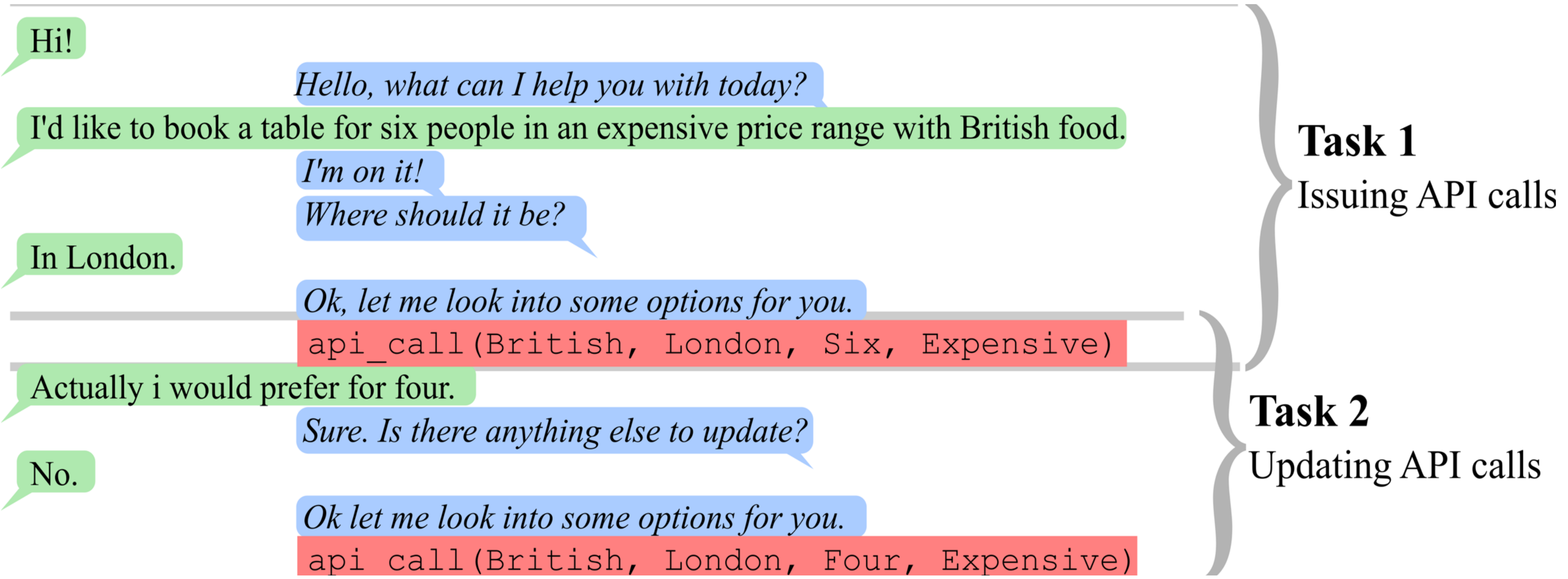
https://fb.ai/babi

https://fb.ai/the-long-game-towards-understanding-dialog/

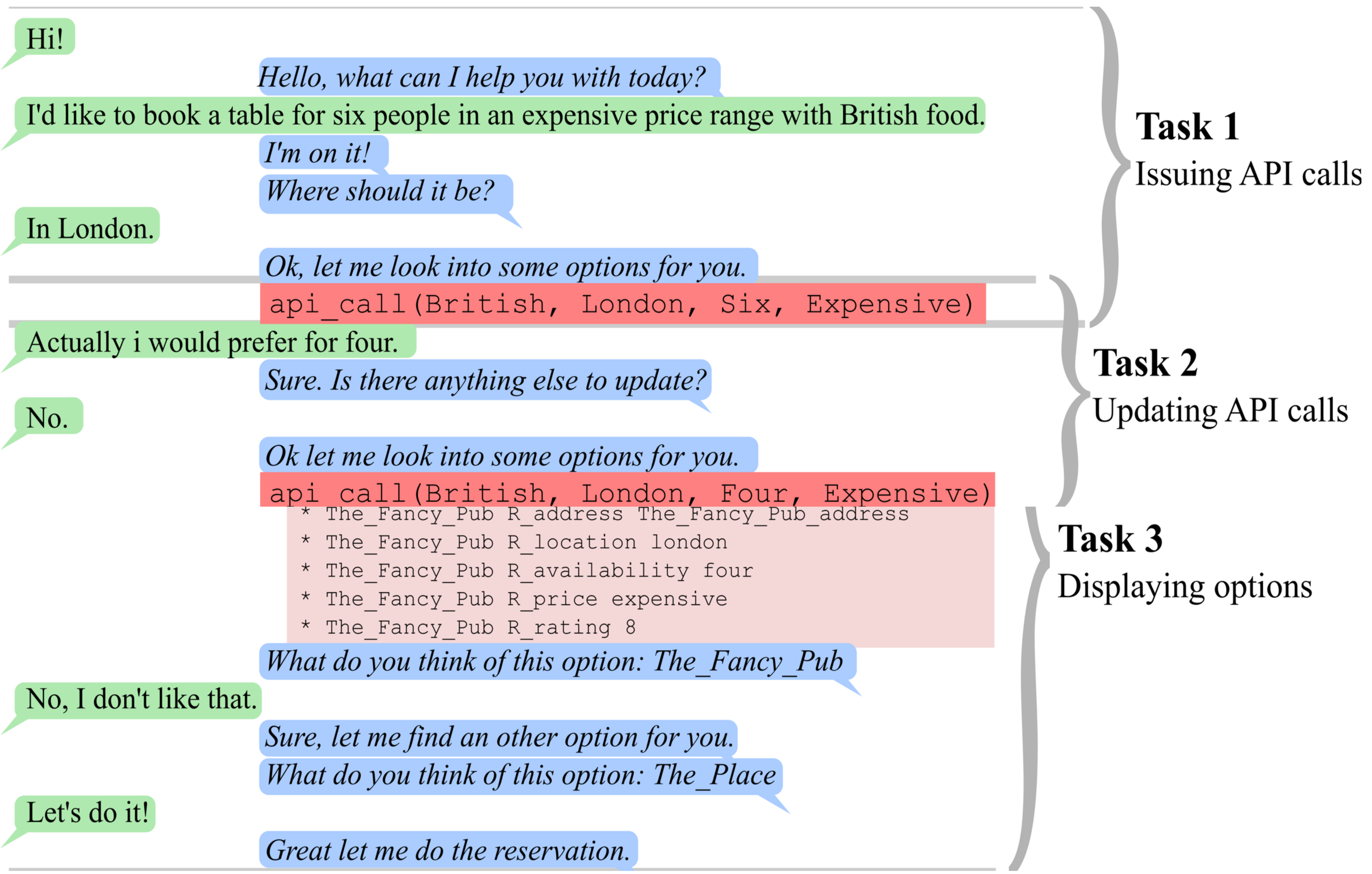
Goal-oriented dialog



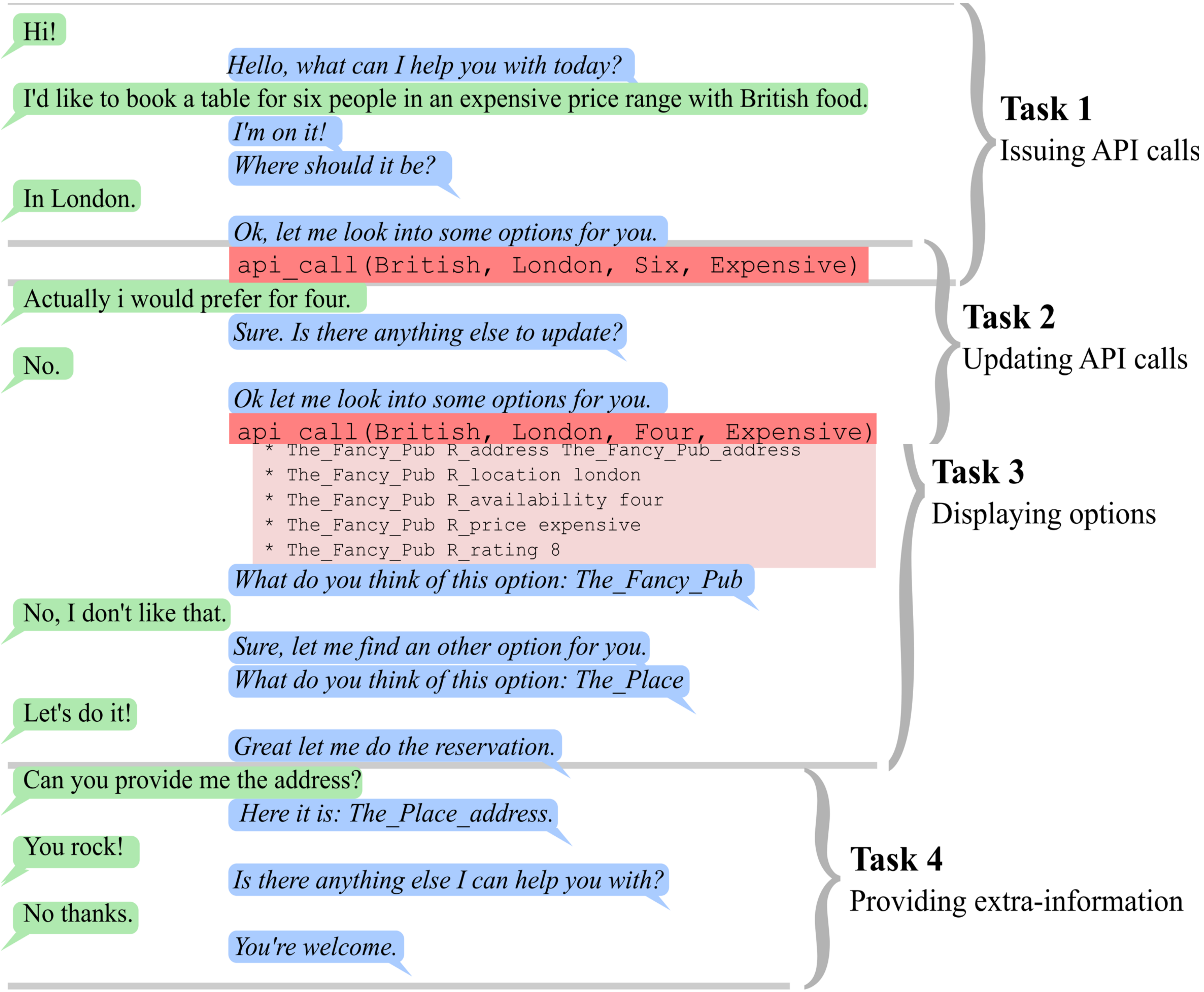
Goal-oriented dialog



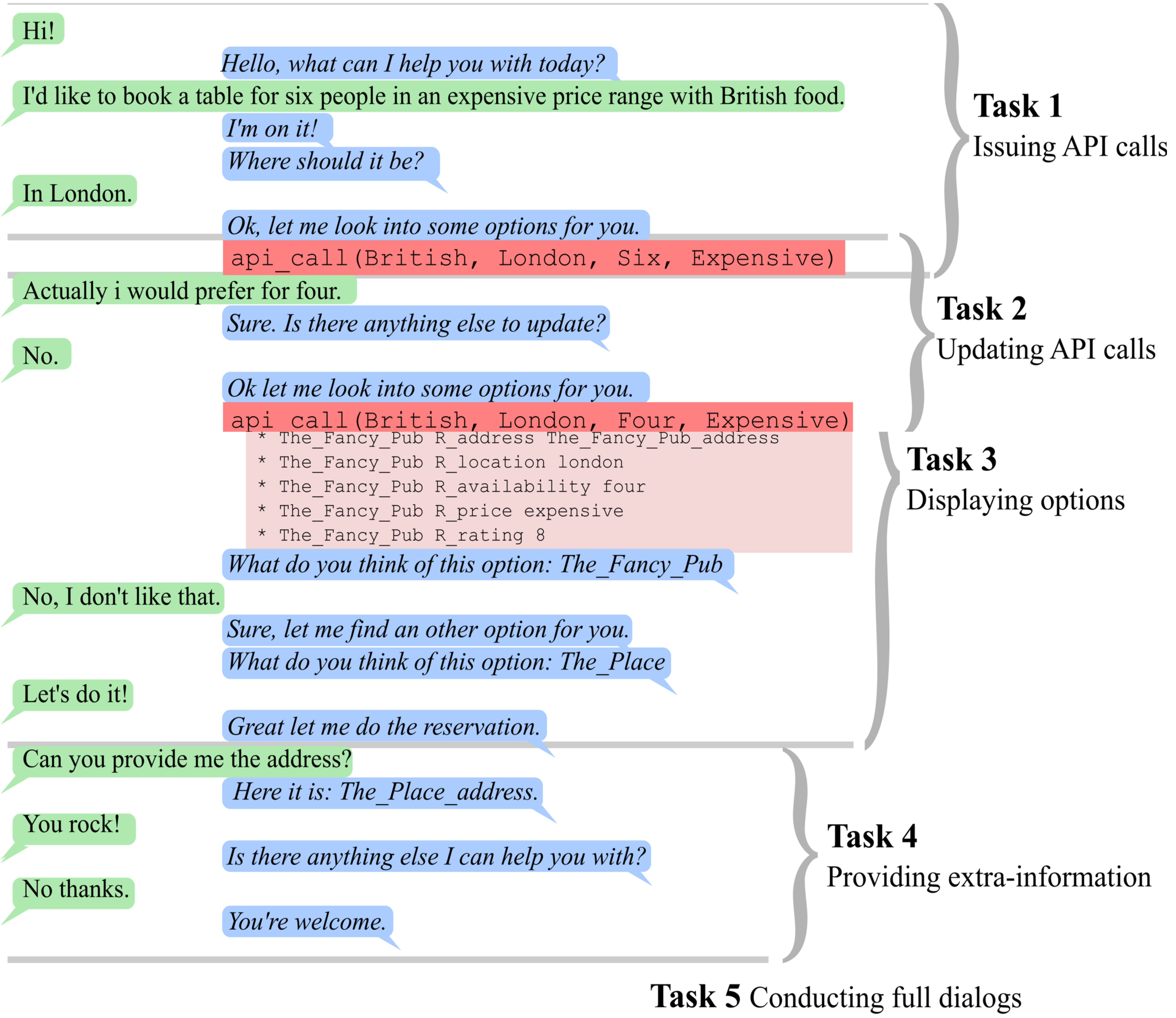
Goal-oriented dialog



Goal-oriented dialog



Goal-oriented dialog



Baselines

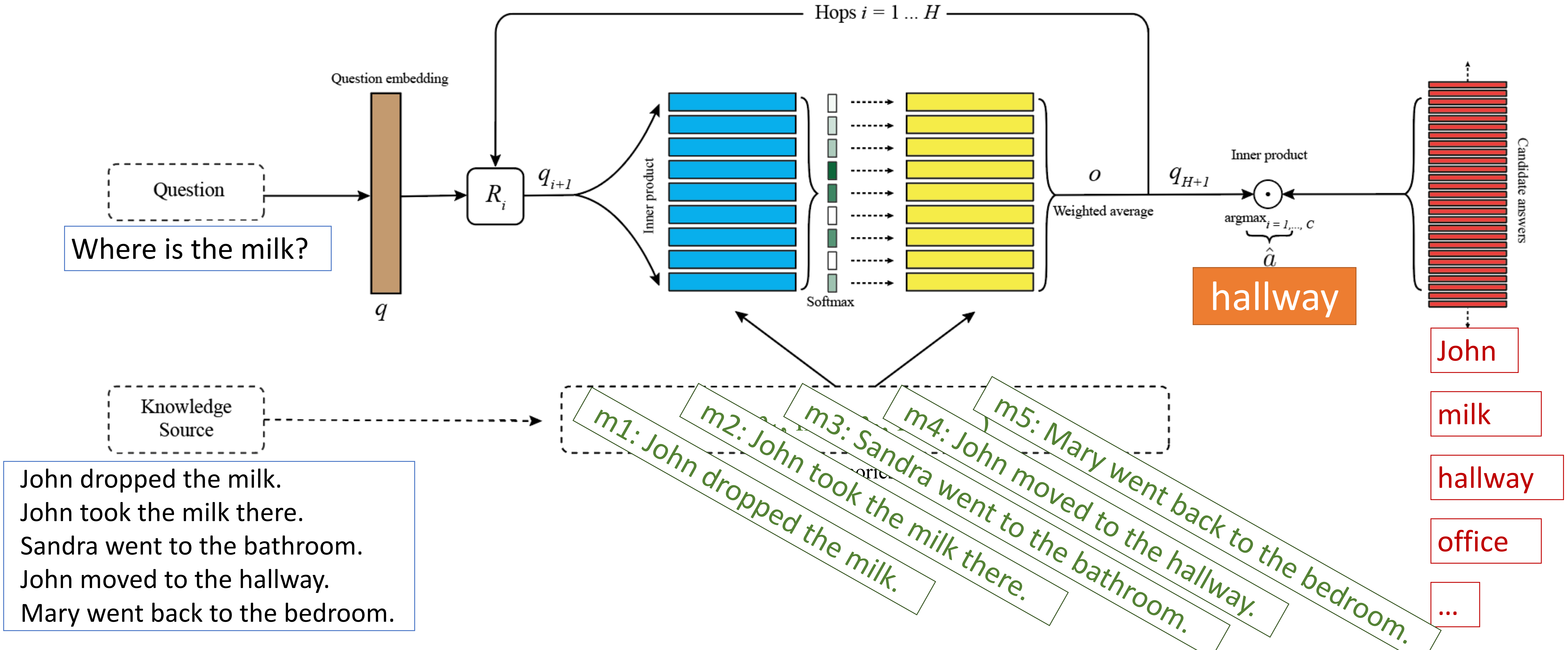
- TF-IDF Match:
  - Matching score between input and candidate
  - Input is either last utterance or all history, whichever is best
- Nearest neighbor:
  - Training: utterance-response
  - Metric: word overlap
- Supervised embeddings: embeddings trained to rank

Testing an end-to-end system: memory network

- Memory networks: combine a large memory with learning component that can read and write to it
- End-to-End version (Sukhbaatar et al, 2015)
  - Soft attention
  - Multiple lookups (hops)
  - End-to-End training with backpropagation
  - Supervision only from final output, not attention

https://fb.ai/the-long-game-towards-understanding-dialog/

Memory Networks (Weston et al., ICLR15; Sukhbaatar et al., NIPS15)



Results

Task	Rule-based Systems	TF-IDF Match no type	TF-IDF Match + type	Nearest Neighbor	Supervised Embeddings	Memory Networks no match type	Memory Networks + match type
T1: Issuing API calls	*100 (100)*	5.6 (0)	22.4 (0)	55.1 (0)	**100** (100)	**99.9** (99.6)	**100** (100)
T2: Updating API calls	*100 (100)*	3.4 (0)	16.4 (0)	68.3 (0)	68.4 (0)	**100** (100)	98.3 (83.9)
T3: Displaying options	*100 (100)*	8.0 (0)	8.0 (0)	58.8 (0)	64.9 (0)	**74.9** (2.0)	**74.9** (0)
T4: Providing information	*100 (100)*	9.5 (0)	17.8 (0)	28.6 (0)	57.2 (0)	59.5 (3.0)	**100** (100)
T5: Full dialogs	*100 (100)*	4.6 (0)	8.1 (0)	57.1 (0)	75.4 (0)	**96.1** (49.4)	93.4 (19.7)
T1(OOV): Issuing API calls	*100 (100)*	5.8 (0)	22.4 (0)	44.1 (0)	60.0 (0)	72.3 (0)	**96.5** (82.7)
T2(OOV): Updating API calls	*100 (100)*	3.5 (0)	16.8 (0)	68.3 (0)	68.3 (0)	78.9 (0)	**94.5** (48.4)
T3(OOV): Displaying options	*100 (100)*	8.3 (0)	8.3 (0)	58.8 (0)	65.0 (0)	74.4 (0)	**75.2** (0)
T4(OOV): Providing inform.	*100 (100)*	9.8 (0)	17.2 (0)	28.6 (0)	57.0 (0)	57.6 (0)	**100** (100)
T5(OOV): Full dialogs	*100 (100)*	4.6 (0)	9.0 (0)	48.4 (0)	58.2 (0)	65.5 (0)	**77.7** (0)
T6: Dialog state tracking 2	33.3 (0)	1.6 (0)	1.6 (0)	21.9 (0)	22.6 (0)	**41.1** (0)	**41.0** (0)
Concierge(*)	n/a	1.1 (0.2)	n/a	13.4 (0.5)	14.6 (0.5)	**16.7** (1.2)	n/a(†)

Results

- Sanity check: rule-based systems achieve 100%
- Information retrieval TF-IDF performs worst
  - This is different from non-goal-directed results
  - Fast pace of goal-directed conversation -> fewer matches?
- Supervised embeddings: better but solve only T1
  - Good at API calls and hello-goodbye
  - Bad at displaying options, providing info, updating calls

Results

- Memory networks outperform IR and supervised embeddings
- Solve T1 and T2 adequately
- Fail a bit less badly at T3 and T4
- Adding matching features makes T4 solvable, not T3
- Same overall pattern of results on real data

MemNN attention on Task 1: good

Attention homes in on the 4 slots

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]**		

MemNN attention on Task 2: good

Attention focuses on first api call and updated slot

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
3	User	may i have a table in paris	.061	.072	.040
4	Bot	i'm on it	.026	.012	.001
5	User	<silence>	.087	.042	.012
6	Bot	any preference on a type of cuisine	.026	.023	.001
7	User	i love indian food	.081	.070	.055
8	Bot	how many people would be in your party	.025	.006	.001
9	User	we will be six	.059	.051	.018
10	Bot	which price range are looking for	.038	.043	.004
11	User	in a moderate price range please	.080	.095	.096
12	Bot	ok let me look into some options for you	.025	.042	.003
13	User	<silence>	.127	.069	.032
14	Bot	`api_call indian paris six moderate`	.062	.113	.043
15	User	instead could it be with french food	.188	.311	.683
16	Bot	sure is there anything else to update	.016	.007	.001
17	User	no	.028	.013	.007
18	Bot	ok let me look into some options for you	.011	.006	.000
User input		<silence>			
Correct answer		`api_call french paris six moderate`			
Predicted answer		`api_call french paris six moderate` **[Correct]**			

MemNN attention on Task 3: fail

Not attending rating

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
14	Bot	`api_call indian paris six moderate`	.012	.000	.000
15	User	instead could it be with french food	.067	.103	.147
20	Bot	`api_call french paris six moderate`	.012	.000	.000
21	User	`resto_1 r_phone rest_1_phone`	.018	.004	.000
23	User	`resto_1 r_cuisine french`	.029	.005	.000
24	User	`resto_1 r_location paris`	.060	.292	.094
25	User	`resto_1 r_number six`	.050	.298	.745
26	User	`resto_1 r_price moderate`	.060	.090	.002
27	User	`resto_1 r_rating 6`	.016	.002	.000
30	User	`resto_2 r_cuisine french`	.031	.007	.000
31	User	`resto_2 r_location paris`	.040	.081	.004
32	User	`resto_2 r_number six`	.020	.012	.000
33	User	`resto_2 r_price moderate`	.029	.009	.000
37	User	`resto_3 r_cuisine french`	.014	.001	.000
38	User	`resto_3 r_location paris`	.028	.016	.001
39	User	`resto_3 r_number six`	.024	.022	.004
40	User	`resto_3 r_price moderate`	.039	.015	.001
User input		<silence>			
Correct answer		what do you think of this option: resto_1			
Predicted answer		what do you think of this option: resto_1	[Correct]		

MemNN attention on Task 4: fail

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
14	Bot	api_call indian paris six moderate	.006	.000	.000
15	User	instead could it be with french food	.024	.011	.007
20	Bot	api_call french paris six moderate	.005	.000	.001
21	User	resto_1 r_phone resto_1_phone	.011	.005	.004
22	User	resto_1 r_address resto_1_address	.018	.004	.001
23	User	resto_1 r_cuisine french	.018	.003	.001
24	User	resto_1 r_location paris	.068	.091	.108
25	User	resto_1 r_number six	.086	.078	.020
26	User	resto_1 r_price moderate	.070	.225	.369
27	User	resto_1 r_rating 6	.014	.006	.008
28	User	resto_2 r_phone resto_2_phone	.015	.009	.006
29	User	resto_2 r_address resto_2_address	.014	.004	.001
31	User	resto_2 r_location paris	.075	.176	.193
32	User	resto_2 r_number six	.100	.126	.026
33	User	resto_2 r_price moderate	.038	.090	.167
35	User	resto_3 r_phone resto_3_phone	.004	.001	.001
36	User	resto_3 r_address resto_3_address	.005	.002	.001
37	User	resto_3 r_location paris	.028	.028	.026
39	User	resto_3 r_number six	.039	.013	.002
40	User	resto_3 r_price moderate	.018	.008	.013
42	Bot	what do you think of this option: resto_1	.074	.001	.000
43	User	let's do it	.032	.004	.001
44	Bot	great let me do the reservation	.003	.000	.000
User input		do you have its address			
Correct answer		here it is resto_1_address			
Predicted answer		here it is: resto_8_address	**[Incorrect]**		

MemNN attention on Concierge real data

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]	

And now?

- Research moves fast: better results since publication already (e.g., Eric and Manning 2017)
- Harder datasets in the works with more challenging features
- -> They will be a DSTC Track this year, try them ☺ https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/

Thanks!

https://fb.ai/babi

https://fb.ai/the-long-game-towards-understanding-dialog/

Dataset statistics

Table 1: **Data used in this paper.** Tasks 1-5 were generated using our simulator and share the same KB. Task 6 was converted from the 2nd Dialog State Tracking Challenge (Henderson *et al.*, 2014a). *Concierge* is made of chats extracted from a real online concierge service. $^{(*)}$ Tasks 1-5 have two test sets, one using the vocabulary of the training set and the other using out-of-vocabulary words.

	Tasks	T1	T2	T3	T4	T5	T6	Concierge
DIALOGS *Average statistics*	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 *Tasks 1-5 share the same data source*	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

Related work

- Most successful goal-oriented dialog systems: model conversation as POMDP (Young et al 2013)
  - Requires many handcrafted features: hard to generalize
- Existing data (Serban et al 2015):
  - Designed to train components of state tracker(Henderson et al 2014)
  - Not open source or require participation to a challenge
  - Noisy if based on interaction of users with a system

Results

Task	Supervised Embeddings						Memory Networks			
	no match type no bigram		+ match type no bigram		+ bigrams no match type		no match type		+ match type	
T1: Issuing API calls	**100**	(100)	83.2	(0)	98.6	(92.4)	**99.9**	(99.6)	**100**	(100)
T2: Updating API calls	68.4	(0)	68.4	(0)	68.3	(0)	**100**	(100)	98.3	(83.9)
T3: Displaying options	64.9	(0)	64.9	(0)	64.9	(0)	**74.9**	(2.0)	**74.9**	(0)
T4: Providing information	57.2	(0)	57.2	(0)	57.3	(0)	59.5	(3.0)	**100**	(100)
T5: Full dialogs	75.4	(0)	76.2	(0)	83.4	(0)	**96.1**	(49.4)	93.4	(19.7)
T1(OOV): Issuing API calls	60.0	(0)	67.2	(0)	58.8	(0)	72.3	(0)	**96.5**	(82.7)
T2(OOV): Updating API calls	68.3	(0)	68.3	(0)	68.3	(0)	78.9	(0)	**94.5**	(48.4)
T3(OOV): Displaying options	65.0	(0)	65.0	(0)	62.1	(0)	74.4	(0)	**75.2**	(0)
T4(OOV): Providing inform.	57.0	(0)	57.1	(0)	57.0	(0)	57.6	(0)	**100**	(100)
T5(OOV): Full dialogs	58.2	(0)	64.4	(0)	50.4	(0)	65.5	(0)	**77.7**	(0)
T6: Dialog state tracking 2	22.6	(0)	22.1	(0)	21.8	(0)	**41.1**	(0)	**41.0**	(0)