

Online Forum Liu et al. (2018)	A: 求推荐励志电影, 谢谢大家 (Any inspirational movies? Thanks everyone). B: 《当幸福来敲门》 (<i>The Pursuit of Happiness</i>).
Crowd-sourced (Constrained) Zhou et al. (2018b)	A: Hey have you seen the inception? B: No, I have not but have heard of it. What is it about ? A: It's about extractors that perform experiments using...
Crowd-sourced (unconstrained) Zhou et al. (2020)	A: 知道重庆森林这个电影吗 (Do you know the movie <i>Chungking Express</i>)? B: 知道啊, 是王家卫导演的 (Yes, it's directed by Wong Kar-Wai). A: 具体是哪年上映的你还记得吗? (Remember which year it was on)? B: 记得, 是在1994年07月14日 (Yes, July 14th, 1994)...
Ours	A: 威尔史密斯演技真的很棒 (Will Smith's acting skill is really good). B: 他的当幸福来敲门太经典了 (His <i>The Pursuit of Happiness</i> is a classic). A: 一直都挂在电影排行榜靠前的位置 (That's always among top ranked movies). B: 嗯嗯, 这部电影真的很励志啊 (Yes, it's really motivational). A: 威尔史密斯也演出了很惨的感觉了 (Will Smith plays like he is a real tragedy). B: 演技特别好 (Yes, he acts pretty well).

Table 10: Comparison between our dataset and others.

A Dataset Collection

Table 10 shows examples comparing our dataset and the others. As observed, forum conversations are mostly single-turn QA or comments. Current crowd-sourced datasets are either collected on constrained scenarios (the scenario in (Zhou et al., 2018b) fixed the roles in a conversation as one introducer and one listener), or unconstrained but prompting people to deliberately connect knowledge. Our dataset simulates real-life conversations to the largest extent.

We classify the utterances into one of 15 aspects. The definitions, counts, and examples of them are shown in Table 11. When annotating the corpus, tutorials and examples are provided to the annotators, we show some examples of in the following tables. All the examples are provided only in Chinese as that is the native language among annotators.

Aspects	Count(%)	Description	Example
电影名(name)	17.41	Name of a movie.	Its name is so funny
导演(director)	2.44	Director of a movie.	Who played Jack in the movie?
演员(actor)	9.24	Actor of a movie.	I wonder who picked him as the leading actor.
类型(type)	7.12	Type of a movie.	I don't really like horror movies.
角色(role)	10.01	Role of a movie.	I like Bumblebee, he is cute.
区域(region)	2.95	Region of a movie.	Is it a Japanese movie?
上映时间(time)	2.02	Release time of a movie.	I remember it was in 1997, still at high school.
剧情(plot)	24.30	Plot of a movie.	What happened to him?
台词(line)	1.51	Line of a movie.	That line becomes a popular word that year.
奖项(awards)	1.17	Award of a movie.	It won a lot of awards that year.
票房(gross)	1.00	Gross of a movie.	I doubt if its gross will be satisfying.
评分(rating)	4.04	Rating of a movie.	It's clearly the best movie I've watched.
资源(website)	1.25	Website of a movie.	You can watch it on iqiyi.com!
音乐(music)	2.19	Music in the movie.	Its theme song is "my heart will go on".
其他(other)	12.31	Other aspects.	I like her costume so much!

Table 11: Counts, descriptions and examples of the aspects. *Examples are shown only in translated English for space limit.*

B Human Evaluation

As for the four human evaluation metrics. The first two will focus only on the conversational backbones without considering domain knowledge. The second two will check if the responses can provide informative and correct responses powered by domain knowledge. The detailed definitions of them are:

A: 一直有听到《千与千寻》，却未曾真的看过这部电影	告知事实_类型; 告知感受_其他
B: 怎么不去看	其他
A: 话说你看了吗	其他
B: 看了，挺感动的	告知感受_其他
A: 主要讲什么的	请求事实_剧情
B: 一个小女孩，千寻，不畏艰难救了自己的父母和朋友	告知事实_角色; 告知事实_剧情
A: 海报里的小女孩看起来还很小	告知事实_角色
B: 所以啊，觉很她很勇敢	告知感受_其他
A: 听上去好像还挺好看的	告知感受_评分
B: 是啊，你快去看看吧	其他

A: 发现一部有点年代感的电影	告知感受_其他
B: 你说什么类型的	请求事实_类型
A: 特别反映现在的问题女人的地位	告知事实_剧情
B: 啥电影	请求事实_电影名
A: 菊豆，巩俐演的	告知事实_电影名告知事实_演员名
B: 卧槽，她啊实力派啊自带气质真的	告知感受_演员名
A: 对啊我才看完	其他
B: 怎么样	请求感受_剧情
A: 有点点压抑，还是觉得我们现在的生活幸福	告知感受_剧情
B: 哈哈我也去看看	其他

A: 《蚁人》你看了没有	告知事实_电影名
B: 蚁人绝对是最接地气的超级英雄!	告知事实_角色
A: 从小偷变英雄 (因为是没有超能力吗?)	告知事实_其他
B: 好看死了!! 每天炸纽约炸香港炸洛杉矶都看醉了!	告知感受_评分
A: 在玩具房里的打斗戏太有意思了!	告知事实_剧情; 告知感受_评分
B: 笑点也挺多的，看得我很开心。	告知感受_评分
A: Paul不是最帅的但绝对是最有味道的!	告知事实_角色
B: 就是最后十秒钟变星际穿越!真心吐槽	告知事实_剧情
A: 各种官方自带吐槽和自黑恶搞真是入戏不过三秒	告知感受_其他
B: 哈哈，确实是	其他

A: 你有没有看过哆啦A梦的电影吗	告知事实_电影名
B: 你说的是哪一部?!	请求事实_电影名
A: 大雄的月球探险记	告知事实_电影名
B: 对啊，哆啦A梦小时候最喜歡了，口袋里面什么都有	告知事实_剧情
A: 哈哈对，这里面他们去月球找玉兔，作者想象力很丰富!	告知事实_剧情
B: 后半部分其实有点融梗《星球大战》+《阿丽塔》	告知事实_电影名
A: 但也没有看出来有什么违和点啊	告知感受_评分
B: 恩呢，那倒是一直是萌物与科普，以及不变的友情。	告知感受_剧情
A: 最后胖虎把那个东西送给那个小孩的时候就哭了	告知事实_剧情
B: 你也一样嘛?	其他
A: 对啊对啊，就是蓝胖子的口袋东西太多了吧	告知事实_剧情
B: 哈哈当然全都靠道具	告知事实_剧情

1. **Sensibleness:** If the response makes sense given the dialogue context. *Specific domain knowledge is ignored.* For example, if the user asks “Who directed the movie Inception?”, the bot replies “Steven Spielberg! I like him so much!!” The answer is wrong but still considered sensible given the context.
2. **Engagement:** If the response is engaging for the user to follow. The definition is generally similar to diversity, or interestingness metric used for open-domain chitchat. If the bot simply passively replies messages like “That’s good” or “No idea”, it will be marked as non-engaging.
3. **Factuality:** Factuality checks the information correctness in case the model fabricate wrong facts. This is orthogonal to the above metrics. The response can be factual but not sensible or the other way around. The same for Engagement.
4. **Informativeness:** Informativeness checks if the response contains new information about the movie. If the bot replies sth like “That’s so interesting, I’ve never heard of that!”. It is considered specific yet non-informative, because the bot provides no new information.

Intuitively a good closed-domain chatbot should balance well all these skills. Sensibleness plus engagement can evaluate its general performance as a chatbot. Factuality and informativeness can evaluate its capability at actively injecting appropriate domain knowledge into the dialogue.

All the three metrics are evaluated by three crowd-workers each except for factuality. As evaluating factuality requires in-depth knowledge about one movie, it is quite difficult for random human evaluators to judge them. Even if we filter to only keep people who have watched a movie, it is hard to guarantee they can recall all the scenes in the movie. Therefore, the factuality check is only done by the person who performed this dialogue. In the static evaluation, it is evaluated by the annotator who produced the reference response. In the interactive evaluation, it is evaluated by the person who chat with the bot. It is nevertheless not accurate though. However, if the bot can cheat the human into believing its false information, it can also somehow be considered a “success”. We provide examples for guiding the human evaluators in Table 12.

Table 13 shows some interactive examples with humans. We observe Mitsuku can XiaoIce perform decently in single-turn exchanges but strongly struggle at understanding multi-turn user intents. Most conversations stop at turn 4 and will not move on.

Conversation	Score
Input: Do you know who directed the movie Inception? Bot: I know! It was Christopher Nolan! I like him so much!	Sensibleness = 1; Engagement = 1 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: The movie is amazing to imagine about a dream within a dream.	Sensibleness = 0 ; Engagement = 1 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: Christopher Nolan.	Sensibleness = 1; Engagement = 0 Factuality = 1, Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: I know! It was Steven Spielberg! I like him so much!	Sensibleness = 1; Engagement = 1 Factuality = 0 , Informativeness = 1
Input: Do you know who directed the movie Inception? Bot: I don’t remember. Do you like it?	Sensibleness = 1; Engagement = 1 Factuality = 1, Informativeness = 0
Input: Do you know who directed the movie Inception? Bot: Steven Spielberg.	Sensibleness = 1; Engagement = 0 Factuality = 0 , Informativeness = 1

Table 12: Examples for human evaluation scores.

Mistsuku	XiaoIce	Ours
<p><i>User:</i> Do you know who directed the movie Inception?</p> <p><i>Bot:</i> No idea. I will guess at Steven Spielberg.</p> <p><i>User:</i> Maybe, it does look like his style.</p> <p><i>Bot:</i> Interesting comparison.</p>	<p><i>User:</i> 你喜欢看感人的电影吗</p> <p><i>Bot:</i> 我喜欢更加另类一些的</p> <p><i>User:</i> 比如呢</p> <p><i>Bot:</i> 比如?</p>	<p><i>User:</i> 你喜欢看感人的电影吗</p> <p><i>Bot:</i> 喜欢啊, 你有什么推荐吗</p> <p><i>User:</i> 我不是药神很感人</p> <p><i>Bot:</i> 确实是一部很良心的电影</p>
<p><i>User:</i> have you watched any movies recently?</p> <p><i>Bot:</i> My favourite movie is Terminator. Have you seen it?</p> <p><i>User:</i> not really, what is it about?</p> <p><i>Bot:</i> If not that, what?</p>	<p><i>User:</i> 泰坦尼克号主角是谁</p> <p><i>Bot:</i> 这我还真知道, 莱昂纳多·迪卡普里奥主演的</p> <p><i>User:</i> 那是谁导演了呢</p> <p><i>Bot:</i> 你想让我说什么?</p>	<p><i>User:</i> 确实, 徐峥演的也好</p> <p><i>Bot:</i> 徐峥的演绎一直很走心</p> <p><i>User:</i> 不过我最喜欢的还是里面的小黄毛</p> <p><i>Bot:</i> 小黄毛好难啊, 太不容易了</p>
<p><i>User:</i> I like the Avengers series so much</p> <p><i>Bot:</i> Yes. My favourite is Iron Man, as he's a kind of robot. You like Avengers.</p> <p><i>User:</i> Exactly! Can't accept he died in the end, sigh</p> <p><i>Bot:</i> You and I are on the same wavelength. I often wonder if it can.</p>	<p><i>User:</i> 你看过大话西游吗</p> <p><i>Bot:</i> 你可以再说一遍吗</p>	<p><i>User:</i> 给我哭完了</p> <p><i>Bot:</i> 看完你有什么感想呢</p> <p><i>User:</i> 最大的病就是穷病啊</p> <p><i>Bot:</i> 确实是这样, 希望我们能平安</p> <p><i>User:</i> 平安是最大的幸福了</p> <p><i>Bot:</i> 我觉得还行</p>

Table 13: Sampled conversations from the interactive evaluation.