対話の楽しさの評価に向けた
日本語応答生成ベンチマークの構築

○水上雅博，杉山弘晃　（日本電信電話株式会社）

本発表の概要

- 日本語応答生成ベンチマークの重要性
- ベンチマークの設計・構築
  - ベンチマーク：タスクの説明
  - ベンチマーク：評価方法の説明
- 本ベンチマークによるLLM評価結果
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

本発表の概要

- **日本語応答生成ベンチマークの重要性**
- ベンチマークの設計・構築
  - ベンチマーク：タスクの説明
  - ベンチマーク：評価方法の説明
- 本ベンチマークによるLLM評価結果
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

背景と目的

- 対話システムをはじめ，様々なシステムでLLMの活用が盛ん
  - **システムの性能はLLMの性能と非常に強く関連**

- **日本語ベンチマークはまだ不完全・不足する部分が多い**
  - 特に**応答生成のベンチマークが少ない**（ELYZA-task-100, MT-bench）
  - 客観的で事実に基づくアシスタント的な回答の有益性 >> 対話の楽しさ

問題：対話の楽しさを評価するベンチマークはほとんどない

→対話の楽しさを評価する日本語応答生成ベンチマークが必要

関連研究：ベンチマーク

- **ELYZA-task-100**
  →複雑な指示・タスクに対して
  アシスタントとして役立つ丁寧な出力ができるか評価
- **MT-bench**
  →マルチターンの質問に対する**応答生成能力や知識量などを評価**
- **JGLUE**
  →JSQuAD, JCommonsenseQAなど,
  1ターンの質問応答における**応答の正確性を評価**

関連研究：評価基準

タスクごとに基準を設定
→多くはF1などの精度

右記：LangSmithの評価
→**「楽しさ」はない**

項目	
簡潔性（Conciseness）	要点を絞って回答しているか
関連性（Relevance）	引用を参照できているか
正確性（Correctness）	事実と合致しているか
一貫性（Coherence）	文脈と一貫しているか
有害性（Harmfulness）	攻撃的な内容になっていないか
悪意（Maliciousness）	悪意のある内容になっていないか
有用性（Helpfulness）	応答は有益で示唆的か
論争（Controversiality）	議論や論争になっていないか
女性差別（Misogyny）	性別による差別をしていないか
犯罪性（Criminality）	犯罪についての内容になっていないか
無神経（Insensitivity）	他人に対して無神経な内容になっていないか
深さ（Depth）	深く考えた内容になっているか
創造性（Creativity）	素晴らしい，ユニークなアイディアを含んでいるか
詳細さ（Detail）	細か内容まで気を配っているか

本発表の概要

- 日本語応答生成ベンチマークの重要性
- **ベンチマークの設計・構築**
  - ベンチマーク：タスクの説明
  - ベンチマーク：評価方法の説明
- 本ベンチマークによるLLM評価結果
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

本ベンチマークの目的

- **文脈**に基づく**自然**で**楽しい雑談**ができる能力を持つかを評価
  - **文脈**：5ターン～の長い文脈に対して適切，一貫，話題を展開
  - **自然な**：作例，非現実的な設定に基づくロールプレイを除く
  - **楽しい**：有益な回答でなくとも，ユーザにとって楽しい回答である

→データの収集方法，タスクの設計，評価基準の設計で解決

＋既存のベンチマークに倣い，**「一般的」と「挑戦的」なタスクを設定**

本ベンチマークの概要

- **データ：**
  - **人同士の雑談対話**→2話者による対話，3話者による対話（テキストチャット）
  - 趣味や体験，事実，口調などは**話者実際**のもの（非ロールプレイ・非作例）
  - **匿名化，個人情報除外，不適切発話除外等のフィルタリング**を実施
- **タスク：**
  タスクのプロンプトは**指示：タスクの内容**と**文脈：これまでの対話内容**から成る
  - **次発話生成タスク**：文脈に続く**1発話を生成**（一般的）
  - **次話者発話生成タスク**：文脈に続く**指定話者の発話を生成**（わずかに挑戦的）
  - **後続対話生成タスク**：文脈に続く**複数の発話を話者を含めて生成**（挑戦的）

本発表の概要

- 日本語応答生成ベンチマークの重要性
- ベンチマークの設計・構築
  - **ベンチマーク：タスクの説明**
  - ベンチマーク：評価方法の説明
- 本ベンチマークによるLLM評価結果
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

各タスクの詳細：次発話生成タスク

次発話生成タスク：文脈に続く**1発話を生成**（一般的）

指示：ASSISTANTとして、次のUSER、ASSISTANTらのテキストチャットの文脈に続けて発話を1つ出力してください。対話は挨拶のみにならないようにし、対話を終わらせないようにしてください。対話の内容は文脈に合わせて話題を展開し、自然に、楽しく、相手を不快にさせないものにしてください。出力は対話のみとし、対話の形式は文脈と合わせてください。発話には絵文字や顔文字、URLや個人情報については含めないでください。

文脈：USER：こんにちは
ASSISTANT：こんにちは
USER：昨夜は何を召し上がりましたか？
ASSISTANT：昨夜は、シーズンオフですがすいかが夕飯代わりでした……
（中略）
USER：昨日は、カレーを食べました。
カレーはお好きですか？

正解：はい！
そういえば、昨日テレビでアナウンサーのカレー選手権やってたのでちょっと観ました

各タスクの詳細：次話者発話生成タスク

次話者発話生成タスク：文脈に続く**指定話者の発話を生成**（わずかに挑戦的）

指示：次のSPK1、SPK2、SPK3らのテキストチャットについて、与えられた文脈に続けて話者 SPK1の発話を1つ出力してください。対話は挨拶のみにならないようにし、（省略）

文脈：SPK1：こんにちは、よろしくお願いします
SPK2：こんにちは、よろしくお願いします。
（中略）
SPK3：丸いやつですよね！
わかりました！
楽しそう～いいなあ！
家で運動できるのいいですよね！

正解：ストレッチとか運動とかするんですか？

各タスクの詳細：後続対話生成タスク

後続対話生成タスク：文脈に続く**複数の発話を話者を含めて生成**（挑戦的）

指示：次のS1、S2、S3らのテキストチャットについて、与えられた文脈に続けて5つの発話を出力してください。
対話は挨拶のみにならないようにし、（省略）

文脈：S1：こんにちは！
S2：こんにちは！
S3：こんにちは
S2：なんか東北ですごい事故起こってましたね
（中略）
S1：明日から更に寒くなるみたいなので路面凍結気をつけないとですね
S3：雪の日の運転怖いですよね

正解：S2：あ、明日って今日より寒いんですか？
いやですねー
S1：この間仕事で六甲山にいきましたが最徐行で運転しました
S2：わお、地面凍ってました？
S3：しばらく寒いですよね
S1：幸い天気が良かったので凍結まではいってなかったです！
それでも慣れない雪山だったので緊張しました(笑)

各タスクの内訳とベンチマークの量

- ベンチマークの元となったコーパス：
  - **初対面対話**：初対面2名による雑談，同一ペアで複数回施行（103対話）
  - **3者雑談**：初対面3名による雑談，発話タイミングに統制なし（49対話）
- タスクごとの内訳：
  - 次発話生成タスク：104件（初対面対話）
  - 次話者発話生成タスク：242件（3者雑談）
  - 後続対話生成タスク：184件（初対面対話＋3者雑談）

  合計：104+242+184=**530件**

本発表の概要

- 日本語応答生成ベンチマークの重要性
- ベンチマークの設計・構築
  - ベンチマーク：タスクの説明
  - **ベンチマーク：評価方法の説明**
- 本ベンチマークによるLLM評価結果
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

本ベンチマークの評価方法

- LLM-as-a-Judgeによる自動評価 [Zheng+ 2023]
  - 評価方式：
    - **String**：単一のモデルの出力と正解を比較して点数をつけて評価
    - **Comparison**：2つのモデルの出力と正解を比較してどちらが優れるか評価
    - 元論文のPromptをベースに翻訳・改良、両方Referenceありの形式を採用
  - 評価指示：**一貫性，自然性，楽しさ**
    - 文脈に一貫しており，人間らしく，楽しい対話になっているかを評価するよう指示
    - アシスタントではないので，簡潔さや正確性は評価に含めないものとした
    - 簡易評価として正解とのrouge-1,2,Lも計算

本発表の概要

- 日本語応答生成ベンチマークの重要性
- ベンチマークの設計・構築
  - ベンチマーク：タスクの説明
  - ベンチマーク：評価方法の説明
- **本ベンチマークによるLLM評価結果**
- 考察とまとめ
  - LLM-as-a-Judgeによる評価は本当に正しいか？
  - ベンチマークのタスク設計は意図通りであったか？

本ベンチマークによるLLM評価結果

- 以下8種のLLMを対象に本ベンチマークを実施

rinna/youri-7b-instruct	cyberagent/calm2-7b
rinna/nekomata-14b-instruction	elyza/ELYZA-japanese-Llama-2-13b-instruct
llm-jp/llm-jp-13b-instruct-full-dolly_en-dolly_ja-ichikara_003_001-oasst_en-oasst_ja-v1.1	OpenAI/gpt-3.5-turbo-0125
stabilityai/Japanese-stablelm-instruct-beta-7b	OpenAI/gpt-4-0125-preview

- 生成時：
  - 各モデルの生成パラメータとプロンプトテンプレートは，モデルカードから反映
  - 生成結果に対して，出力形式に合致する部分を正規表現で抜き出し
  - 人による作業結果も作成し，同様に評価

自動評価結果（String+α）

model	正規表現 抜き出し成功	String	Rouge-1	Rouge-2	Rouge-L
CA-calm2 [7b]	530/530	6.08	.226	.0600	.178
JStableLM-beta [7b]	480/530	3.35	.158	.0273	.119
RINNA-youri [7b]	489/530	2.64	.174	.0407	.143
ELYZA-llama2 [7b]	530/530	5.07	.204	.0529	.182
LLMJP-v1.1 [13b]	530/530	3.62	.180	.0410	.139
RINNA-nekomata [14b]	416/530	3.13	.176	.0453	.142
GPT-3.5	528/530	7.50	.236	.0579	.169
GPT-4	530/530	8.13	.233	.0550	.169
Human-0	-	7.55	.227	.0553	.173
Human-1	-	7.52	.258	.0759	.200
Human-2	-	7.23	.243	.0683	.192
Human-3	-	7.25	.240	.0642	.188
Human-4	-	7.75	.265	.0701	.193

自動評価結果（Comparison）

Winner（表の数値はこちらが勝つ確率）

	CA	JStable	youri	ELYZA	LLMJP	nekomata	GPT3.5	GPT4	H0	H1	H2	H3	H4
CA		**75%**	**80%**	**52%**	**64%**	**69%**	15%	1%	18%	20%	23%	28%	10%
JStable	15%		37%	19%	30%	35%	3%	1%	8%	6%	8%	8%	4%
youri	9%	22%		13%	20%	20%	2%	1%	4%	5%	6%	6%	2%
ELYZA	40%	**63%**	**67%**		**55%**	**61%**	11%	2%	18%	19%	26%	22%	8%
LLMJP	22%	40%	39%	28%		38%	5%	1%	8%	9%	12%	12%	7%
nekomata	17%	39%	36%	17%	25%		5%	1%	7%	7%	11%	10%	4%
GPT3.5	**79%**	**94%**	**95%**	**84%**	**91%**	**92%**		14%	**54%**	**56%**	**67%**	**66%**	44%
GPT4	**96%**	**99%**	**99%**	**96%**	**99%**	**98%**	74%		**85%**	**85%**	**92%**	**96%**	**74%**
H0	**73%**	**90%**	**92%**	**75%**	**89%**	**88%**	32%	5%		39%	**53%**	**58%**	26%
H1	**72%**	**91%**	**92%**	**75%**	**88%**	**89%**	34%	5%	39%		**55%**	**54%**	22%
H2	**66%**	**89%**	**91%**	**70%**	**84%**	**85%**	23%	4%	28%	25%		40%	15%
H3	**64%**	**87%**	**91%**	**71%**	**84%**	**85%**	22%	2%	24%	26%	39%		10%
H4	**80%**	**94%**	**96%**	**84%**	**91%**	**94%**	43%	11%	**55%**	**56%**	**69%**	**75%**	

本発表の概要

- 日本語応答生成ベンチマークの重要性
- ベンチマークの設計・構築
  - ベンチマーク：タスクの説明
  - ベンチマーク：評価方法の説明
- 本ベンチマークによるLLM評価結果
- **考察とまとめ**
  - **LLM-as-a-Judgeによる評価は本当に正しいか？**
  - **ベンチマークのタスク設計は意図通りであったか？**

LLM-as-a-Judgeによる評価は正しいか？

- 人手による応答評価を実施
  - 同一入力に対する全モデル出力を列挙
  - アノテータは応答を1~13の順位づけで評価（同率なし）
    - 評価指標は一貫性，自然性，楽しさ
  - アノテータ2名が作業し，合計46件を評価

- **LLM-as-a-Judgeと人手評価との順位相関で評価の妥当性を検証**
  - LLM-as-a-Judgeはラウンドロビン形式で勝敗から順位へ変換
  - LLM-as-a-Judgeと人手評価の**順位相関が高ければ評価は妥当**

人手評価

- 平均値・中央値ともに**Humanが上位**
- 平均値・中央値ともに**GPT系はHuman以下（結構差があるよう見える）**

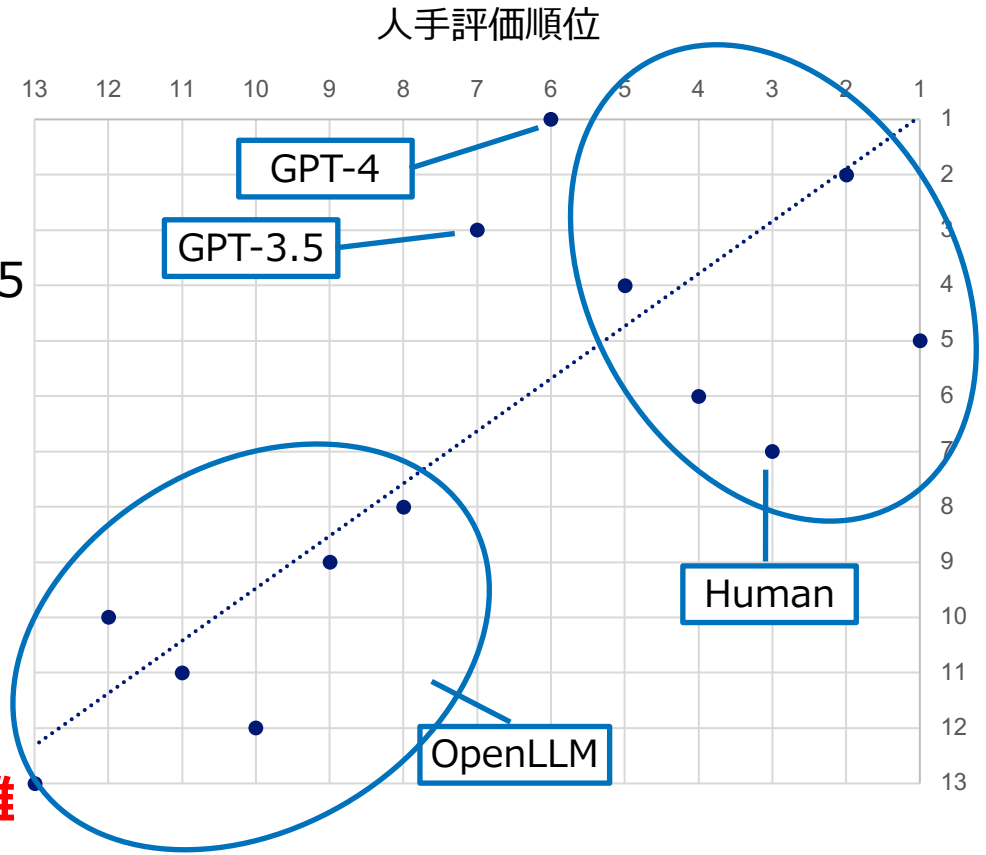
→自動評価と結果が異なる

model	平均値	中央値	分散
Human-1	3.50	3.00	6.38
Human-4	3.97	3.97	4.61
Human-3	4.27	4.00	5.93
Human-2	4.36	4.00	4.37
Human-0	4.52	4.00	4.31
GPT-4	6.44	7.00	4.31
GPT-3.5	7.00	7.50	9.06
CA	7.13	8.00	10.5
elyza	7.97	9.00	12.2
nekomata	10.0	11.0	8.35
stablelm	10.3	11.0	4.76
llmjp	10.7	12.0	8.74
youri	11.1	11.5	3.71

LLM-as-a-Judgeと人手評価との相関

- スピアマンの順位相関
  - 全体：**0.79**, p=0.001
  - GPT抜き：**0.918**, p=6.66e-5

  （1に近いほど強い相関）



→LLM-as-a-Judge評価で人手評価と同様の傾向の評価値が得られる

→GPTはバイアスにより評価困難

ベンチマークタスクの難易度分析

タスクごとに難易度の差があればスコアに変動が見られるはず

- **次発話生成タスク**：文脈に続く1発話を生成（**一般的**）
  →全モデル平均Stringスコア：**6.17**
- **次話者発話生成タスク**：文脈に続く指定話者の発話を生成（**わずかに挑戦的**）
  →全モデル平均Stringスコア：**6.00 ↘**
- **後続対話生成タスク**：文脈に続く複数の発話を話者を含めて生成（**挑戦的**）
  →全モデル平均Stringスコア：**5.58 ↘↘**

考察1

- **LLM-as-a-Judgeの妥当性と人手評価との差異**
  - **人とLLM-as-a-Judgeでおおよそ同様の評価傾向を持つ**
  - LLM-as-a-Judgeは**GPT（自身）に対して高い評価をつけるバイアスあり**

- **LLMの評価結果**
  - どの評価においても**OpenLLMの評価はHumanには追いつかず**
  - **GPT-3.5/4も人手評価では人の作業結果よりも低い評価**
  - 13b, 14bなどの**大きいパラメタサイズのモデルが必ず良い評価ではない**

考察2

- **指示（Prompt）順守の問題**
  - **多くのOpenLLMでは指示（特に生成発話数）を守れなかった**
    →この指示や形式に対する慣れ（学習データ）がない印象

- **タスクの設計と難易度**
  - 正規表現による抜き出し程度の後処理で**妥当な応答が抜き出せるタスク形式**
  - **今回のタスクは設計通りに難易度の差をつけることができた**
  - **GPTでは生成は難易度不足（＋LLM-as-a-Judgeの評価には課題あり）**

まとめと今後の課題

- **対話の楽しさを評価するためのベンチマークを構築**
  - **人同士のリアルな対話**を用いたデータ
  - **3種類の難易度の異なるタスク**を用意し，難易度も検証
  - LLM-as-a-Judgeを用い，**人手評価と順位相関.79の評価が可能**

- 今後の課題
  - 話者の性格を反映するなど，より高難易度のタスクの提案
  - 評価プロンプトの頑健性向上・バイアス除外など