LLM-as-a-judgeへの事後アノテーションによる人補正

Estimating Human Judge Scores through Post-Hoc Annotation of LLM-as-a-Judge

手嶋 毅志 *1　篠塚 健太 *2　松岡 佑知 *1
Takeshi Teshima　Kenta Shinotsuka　Yuchi Matsuoka

*1株式会社リクルート
Recruit Co., Ltd.

The purpose of this study is to enhance chatbot performance evaluation using a Large Language Model (LLM) as a judge through post-hoc annotations. This research addresses the challenge of evaluating chatbot responses, which typically requires significant human effort due to the complexity of natural language data. By implementing post-hoc annotations on LLM judgments, we aim to estimate scores that would be obtained through full human evaluation. Our contribution extends beyond evaluating the biases of LLM-as-a-judge: we propose constructing justifiable estimators of human evaluation results based on LLM judgments. Using a set of acceptable assumptions, we derive a version of the estimator based on post-hoc annotations of the LLM's evaluation labels. We implemented the estimator in an industrial context and experimented with coordinating domain experts. The experiments confirmed the practical applicability of our approach, paving the way for more efficient evaluation processes in AI-powered customer service systems.

1. はじめに

リクルートでは，事業者の業務や経営を支援するための多様なサービスおよびプロダクトを提供している．これらのサービスの利用者数が増加する中で，問い合わせ窓口の効率的なスケールアップが求められており，AI チャットボットによる自動応対の導入が進められている．

チャットボットの応対品質を向上させるためには，評価指標の計測とモニタリングが不可欠である．しかしながら，自然言語データの特性により，従来の評価方法では人手による継続的な評価が必要となり，結果として高コストとなることが課題として挙げられる．また，評価対象データの量によっては複数の評価者が分担することになり，評価基準や品質の統一という課題が増えることになる [Monarch 21]．

このような背景を受けて，大規模言語モデル（LLM）を用いた自動評価手法「LLM-as-a-judge」が最近提案され，普及が進んでいる．この手法は，自然言語による応対を大規模に自動評価しうる有望なアプローチだが，LLM の評価と人間の評価との間には依然として無視できない乖離が存在することが知られている [Zheng 23]．

そこで本発表では，チャットボットの性能評価のための LLM-as-a-judge の適用事例を紹介するとともに，新たな工夫として，LLM の判定結果に対して事後的なアノテーションを行うことにより，人が全量評価した場合に得られるであろう評価指標値を推定することを提案し，その試みについても報告する．

2. 問題設定

本稿では，チャットボットの性能評価における LLM-as-a-judge を検討する．まず本節で，その問題設定と定式化を行う．

2.1 問題設定

LLM-as-a-judge によるチャットボットの性能評価を考える（以下では評価者となる LLM を，ジャッジ LLM と呼称する）．ジャッジ LLM が評価を行う対象は，サービスの利用者とチャットボットの対話を記録したスナップショット（以下では評価ケースと呼称する）である．

本稿では一往復の会話に対する評価のみを考える．一般にチャットボットの対話は複数往復にわたって行われるが，その場合への本稿の内容の拡張は今後の研究課題とする．また，本稿では一回のユーザー入力毎に，チャットボットが複数の回答を返すシステムを考える．例えば，情報検索で拡張した生成（retrieval-augmented generation, RAG; [Lewis 20] など）において，上位の検索結果のそれぞれから複数の回答を生成して利用者に返却する場合がこれに相当する．なおこれは，単答形式の対話も特別な場合として含む．次節にてこれらを定式化する．

2.2 定式化

ユーザーの入力 X に対するチャットボットの回答は，サイズ $K \in \mathbb{N}$ の順序付きタプル $\mathbf{R} = (R^{(1)}, \ldots, R^{(K)})$ であるとし，組 $(X, \mathbf{R})$ を評価ケースと呼ぶ．これら $X, R^{(1)}, \ldots, R^{(K)}$ はいずれも，適当な長さを上限とする文字列の空間に値をとる確率変数である．

続いて，人間やジャッジ LLM に対して，個別の問答対 $(X, R^{(k)})$ を渡すことで，次の $\{0,1\}$-値の評価ラベルを得られるとする．

$$Y^{(k)} := \mathbf{1}\{X \text{ に対して } R^{(k)} \text{ が良い回答と人間が判断 }\}$$

$$Z^{(k)} := \mathbf{1}\{X \text{ に対して } R^{(k)} \text{ が良い回答と LLM が判断 }\}$$

これらは一般に，$(X, R^{(k)})$ を条件付けても定数とは限らない確率変数とする．また，$(X, R^{(k)})$ を条件付けたもとで $Y^{(k)}$ は他と独立（$Z^{(k)}$ についても同様）とする．記号として，これらをまとめた $\mathbf{Y} = (Y^{(k)})_{k=1}^{K}, \mathbf{Z} = (Z^{(k)})_{k=1}^{K}$ を定義する．

これらの個別の問答対 $(X, R^{(k)})$ に対するラベルから，評価ケース $(X, \mathbf{R})$（複数回答を総合した対）に対するラベルを，それぞれ $\max_k Y^{(k)}$（人間の場合）と $\max_k Z^{(k)}$（ジャッジ LLM の場合）により定義する．即ち，1 つの入力ごとにチャットボットが返す複数の回答のうち，いずれかが良い回答と判断されれ

連絡先: 手嶋毅志（株式会社リクルート）
takeshi_teshima@r.recruit.co.jp
https://takeshi-teshima.info/

ば，その複数回答の全体としても良いとするラベルを付ける．

いま，これらの確率変数の組 $(X, \mathbf{R}, \mathbf{Y}, \mathbf{Z})$ について，n 個の独立同分布（i.i.d.）な複製を考える．

$$\{(X_i, \mathbf{R}_i, \mathbf{Y}_i, \mathbf{Z}_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(x, \mathbf{r}, \mathbf{y}, \mathbf{z})$$

このうち，我々はデータセットとして $\{(X_i, \mathbf{R}_i, \mathbf{Z}_i)\}_{i=1}^{n}$ を得ることができるとする．

この定式化のもとで，LLM-as-a-judge に基づくチャットボットの性能評価指標は，ジャッジ LLM により良いと判断された評価ケースの割合

$$\hat{\mu}_Z := \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le k \le K} Z_i^{(k)}$$

で定義される．

既存研究においては，LLM-as-a-judge の信頼性，即ち，一貫性，バイアスの低減，文脈への追従性，などの $Z^{(k)}$ に内在する傾向性（あるいは評価ケース $(X, \mathbf{R})$ との関係における傾向性）の評価，ならびに，人間の評価との一致率，即ち $Y^{(k)} = Z^{(k)}$ の確率などの評価が主要な課題として扱われている [Zheng 23]．

3. 事後アノテーションによる人補正

本節では，ジャッジ LLM の信頼性に関する評価を主課題とするのではなく，ジャッジ LLM の判定結果に対して事後的なアノテーションを行うことにより，人が仮に全量評価を行なっていれば得られるであろう評価指標値を推定することを，研究課題とすることを提案する．

3.1 訂正された目標

以降，K 次元の零ベクトルを $\mathbf{0}$ で表す．本稿で考える評価問題は，以下のように定式化できる．

問題 1. 小節 *2.2*の定式化のもとで，追加的な仮定を置くことは許容しながらも，人間の評価に基づくチャットボットの性能評価指標 $\mu_Y := p\left(\max_{1 \le k \le K} Y^{(k)} = 1\right)$ を推定せよ．

3.2 人間の評価に基づく性能評価指標の表現の導出

問題 1に対する解決策を導出するために，いくつかの仮定を置く．

仮定 1 (独立な偽陽性)**.** 人間による評価結果 $\mathbf{Y}$ が全て「良くない回答」という判断の場合，ジャッジ *LLM* による評価 $Z^{(1)}, \ldots, Z^{(K)}$ は，$Y^{(1)}, \ldots, Y^{(K)}$ に独立にラベルが反転されたものとして振る舞うと仮定する．即ち，

$$p(\mathbf{Z} | \mathbf{Y} = \mathbf{0}) = \prod_{k=1}^{K} p(Z^{(k)} | Y^{(k)} = 0)$$

と仮定する．

仮定 2 ($\mathbf{0}$ に対する従属性の一致)**.** 回答を良くないと判断する事象に関して，人間の場合におけるその事象の非独立性の度合いを，ジャッジ *LLM* がよく捉えられていると仮定する．

$$\frac{p(Z^{(1)} = 0, \ldots, Z^{(K)} = 0)}{p(Z^{(1)} = 0) \cdots p(Z^{(K)} = 0)} = \frac{p(Y^{(1)} = 0, \ldots, Y^{(K)} = 0)}{p(Y^{(1)} = 0) \cdots p(Y^{(K)} = 0)}$$

これは強い仮定であり，この仮定を緩めることやその妥当性の定量的評価は今後の研究課題の一つである．この仮定の妥当性は，使用している LLM の性能に依存する．完全に人間と同じ評価をするジャッジ LLM，即ち，確率変数として $\mathbf{Z} = \mathbf{Y}$ を誘導するジャッジ LLM であれば仮定 2は当然成立することから，この仮定はジャッジ LLM の評価者としての性能改善によって満たされやすくなると期待される．

命題 1 (アラインメントによる人補正指標)**.** 仮定 *1*と仮定 *2*のもとで，以下が成立する．

$$\frac{1 - \mu_Y}{p(\mathbf{Z} = \mathbf{0})} = \sum_{\mathbf{z} \in \{0,1\}^K} \left(\prod_{k=1}^{K} \frac{p(Y^{(k)} = 0 | Z^{(k)}) \cdot p(Z^{(k)})}{p(Z^{(k)} = 0)} \right)$$

証明の概略（命題 *1*）．まず，$1 - \mu_Y = p(\mathbf{Y} = \mathbf{0})$ であるから，方針として，$p(\mathbf{Y} = \mathbf{0}, \mathbf{Z})$ の $\mathbf{Z}$ に関する和をとることを考える．仮定より，以下の式変形が成立する．

$$\begin{aligned}
&p(\mathbf{Y} = \mathbf{0}, \mathbf{Z}) \\
&= p(\mathbf{Z} | \mathbf{Y} = \mathbf{0}) \cdot p(\mathbf{Y} = \mathbf{0}) \\
&= \left(\prod_{k=1}^{K} p(Z^{(k)} | Y^{(k)} = 0) \right) \cdot p(\mathbf{Y} = \mathbf{0}) \quad (\because \text{仮定 1}) \\
&= \left(\prod_{k=1}^{K} \frac{p(Y^{(k)} = 0 | Z^{(k)}) \cdot p(Z^{(k)})}{p(Y^{(k)} = 0)} \right) \cdot p(\mathbf{Y} = \mathbf{0}) \\
&= \left(\prod_{k=1}^{K} p(Y^{(k)} = 0 | Z^{(k)}) \cdot p(Z^{(k)}) \right) \cdot \frac{p(\mathbf{Y} = \mathbf{0})}{\prod_{k=1}^{K} p(Y^{(k)} = 0)} \\
&= \left(\prod_{k=1}^{K} p(Y^{(k)} = 0 | Z^{(k)}) \cdot p(Z^{(k)}) \right) \cdot \frac{p(\mathbf{Z} = \mathbf{0})}{\prod_{k=1}^{K} p(Z^{(k)} = 0)} \\
&\qquad (\because \text{仮定 2})
\end{aligned}$$

よって主張は示された． □

3.3 事後アノテーションによる補正

命題 1は，問題 1に対する解を一つ与えている．即ち，$p(Y^{(k)} = 0 | Z^{(k)}), p(Z^{(k)}), p(\mathbf{Z} = \mathbf{0})$ のそれぞれに推定値を代入することにより，μ_Y の妥当な推定値を得ることができる．

具体的には，$p(Y^{(k)} = 0 | Z^{(k)})$ の推定値は，データセットから $Z^{(k)} = 0$ であるレコードと $Z^{(k)} = 1$ であるレコードをそれぞれ部分サンプリングし，人手によるアノテーションを行うことにより推定できる．事後アノテーションにおける部分サンプリングは，人手でのアノテーションのための予算が許す範囲の個数を，非復元抽出すればよい．

また，$p(Z^{(k)})$ と $p(\mathbf{Z} = \mathbf{0})$ は，所与のデータセットにおけるモデルの出力比率から，それぞれ $\frac{1}{n} \sum_{i=1}^{n} Z_i^{(k)}$ と $\frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{K} (1 - Z_i^{(k)})$ などにより推定できる．

4. 実践

本研究では，実際にチャットボットの（非公開の）評価用データセットにおいて，LLM-as-a-judge とドメインエキスパートによる事後アノテーションを行い，本稿の性能評価指標の推定量による評価を実施した．

事後アノテーションは，具体的には，まず $p(Y^{(k)} = 0 | Z^{(k)})$ が $(k, Z^{(k)})$ によらず一定であることを仮定し，その上で事後アノテーションのために $Z^{(k)} = 0$ と $Z^{(k)} = 1$ のそれぞれの

部分データセットから抽出したサンプルに対し，人手でのアノテーションを行うことで，$p(Y^{(k)}=0|Z^{(k)})$ を推定することとした．また，実際にはアノテーション基準を統一するにあたって，判断基準が曖昧になるボーダーライン上の評価ケースを扱う際の認知負荷を軽減する目的で，$(X, R^{(k)})$ に対する直接のアノテーションではなく，$(X, R^{(k)}, Z^{(k)})$ をアノテーターに提供した上で $Y^{(k)}=Z^{(k)}$ であるか否か，即ち，$Z^{(k)}$ としたジャッジ LLM の判断を受容できるか否かによってアノテーション作業を行った．[*1]

この方法で得た $p(Y^{(k)}=0|Z^{(k)})$ の点推定値（$Z^{(k)}=0,1$）と，小節 3.3で説明した推定量によりジャッジ LLM による評価ラベルを集計して得た $p(Z^{(k)})$ と $p(\mathbf{Z}=\mathbf{0})$ の点推定値を，命題 1の式に代入することで，μ_Y の推定値を得た．

実際に，$Z_i^{(k)}=0$ と $Z_i^{(k)}=1$ のレコードそれぞれを $m=100$ 件ずつ非復元抽出してアノテーションを行った結果，推定値として $p(Y^{(k)}=0|Z^{(k)}=0)\simeq 0.80, p(Y^{(k)}=1|Z^{(k)}=1)\simeq 0.71$ が得られた．

以下は実際の事業上の数値とは無関係なものだが，仮に $K=2$ として，上記以外の推定値がそれぞれ $p(Z^{(k)}=1)\simeq 0.95$（$k\in\{1,\ldots,2\}$），$p(\mathbf{Z}=\mathbf{0})\simeq 0.005$ とすれば，人が全量評価した場合の評価指標値への推定値は $\mu_Y\simeq 0.801$ 程度となる．つまり，ジャッジ LLM の判断をそのまま評価指標として用いると $(1-0.005)=0.995$ 程度のチャットボット性能という評価になるが，人間とジャッジ LLM の間のアラインメント情報による補正を行った評価指標は，0.80 程度となる．このように人補正を行うことで，改善可能な余地がどの程度あるかを定量的に評価することができるという点が，本稿の主張である．

5. まとめ

本研究では，LLM-as-a-judge を用いたチャットボットの性能評価に関する新たなアプローチを提案し，事後的なアノテーションによって人間評価に基づく指標を推定する手法を示した．今後の研究課題として，以下の点が挙げられる．

新たな推定量の構成：本稿で提案した推定量以外にも，どのような仮定のもとで正当化可能な推定量が構成できるかを探索することが重要である．これにより，異なる評価シナリオやデータセットに対して，より柔軟で適用可能な手法の開発が期待される．また，人手による事後アノテーションのサンプルサイズは比較的小さくなることから，実際には事後アノテーションの推定誤差を考慮した区間推定量などを構成して用いることも検討の価値がある．

仮定の緩和：本研究で採用した仮定を弱めることは，評価の信頼性を高めるために重要である．特に，仮定 2を等号制約ではなく定量的な差異の制約とした仮定のもとで導出される推定量や，部分識別に基づく推定量など，弱い仮定から導出できる広く適用可能な推定量を構築することは実用的意義が大きい．

理論的な検証：提案手法の理論的基盤を強化するために，無作為抽出された評価ケースに対するアノテーションを行った場合の人間評価値の推定との比較研究を行うことが考えられる．仮に今回の提案手法で用いるのと同じだけの人的コストを，無作為抽出した評価ケースに対するアノテーションに掛ければ，その単純な集計によっても人間による評価値の推定量を構成することが可能である．この方法で構成する推定量と比べて LLM-as-a-judge を経由する評価値が理論的な見地から良いものとなる条件を明らかにしたい．この比較を通じて，提案手法の優位性や限界をより明確にすることが可能であると期待される．

分布変化への頑健性と適応性：チャットボットの変更や新たなデータの導入に伴う分布の変化に対して，提案手法の頑健性を検討することも重要な研究課題である．これにより，実用的なシナリオにおいて信頼性の高い評価が可能となる．また，それらの分布変化に対して，効率よく追従することができるオンラインな推定量の構成も重要な課題である．

LLM の性能改善による影響：LLM の性能向上が提案手法に与える影響を評価することも重要である．LLM の評価者としての性能を定期的に評価し，その改善が提案手法に及ぼす影響を分析することで，より効果的な評価手法の構築が期待される．

以上の研究課題に取り組むことで，LLM を用いたチャットボットの評価手法がより広範な応用に適したものとなり，実際のビジネス環境での有用性が向上することが期待される．

謝辞

大場夢子さん，近藤都さん，江村航大さん，竹内優太さん，松村尚明さん，上津将士さんをはじめとする，ドメインエキスパートとして実験にご協力いただいた皆様には，アノテーション作業のために貴重な時間を割いていただきました．この場を借りて感謝申し上げます．ありがとうございました．

参考文献

[Lewis 20] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS'20, Red Hook, NY, USA (2020), Curran Associates Inc.

[Monarch 21] Monarch, R. M.: *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*, Manning Publications, Nueva York (2021)

[Zheng 23] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2023)

*1 なお，この際のアノテーション作業工数は，平均で 1 件あたり 1.05 分であった．