QA タスク回答中の趣旨・補足に対する不適・不足検出への LLM 適用

飯田頌平　林友超　宍戸里絵
弥生株式会社 AI・データ戦略部 R&D チーム

概要

本論文では質問応答タスクにおける回答の「趣旨」「補足」を分類し，参照回答との「不適」「不足」を考慮する指標について，その評価を LLM を利用することで自動化する手法について提案する．ここで「趣旨」とは主張の中心となる情報，「補足」はそれを補う情報を指し，さらに出力回答の「趣旨」や「補足」が参照回答にはない情報を述べている場合に「不適」，情報が欠けている場合に「不足」と分類する．文献 [14] で提案されたこの指標について，本論文では自動評価する手法を確立することで，より大規模に評価可能な枠組みを実現可能とした．

1　はじめに

本論文では質問応答タスクにおいて，回答中の「趣旨」「補足」情報まで考慮した評価を行う．さらに出力回答が参照回答と比較して過剰に書かれている箇所や欠けている箇所をそれぞれ「不適」「不足」と定義し，不適度や不足度を計測して出力回答を詳細に分類している．

これらの評価指標は文献 [14] で提案されたものであるが，文献 [14] では「趣旨」「補足」を考慮した出力回答中の「不適」「不足」検出を人手評価によって実施していたため，大規模な評価が出来ないという課題があった．そこで本論文では，前述の手順を人手ではなく LLM を用いて自動化した手法を提案する．提案手法ではまず参照回答の文字列を LLM によって意味単位で「趣旨」「補足」に分類する．次に出力回答の文字列を同様に「趣旨」「補足」へ分類した後，参照回答の各「趣旨」「補足」に対して「不適」「不足」である箇所を LLM により検出する．

さらに LLM による検出結果を文献 [14] で実施された人手評価による検出結果と比較することで，本論文による自動評価の枠組みが人手による評価結果と大きな差異がないことを確認し，文献 [14] による評価指標をより大規模に適用可能なものとした．

2　関連研究

質問応答タスクでは LLM により回答を生成する様々な研究 [4, 3, 13] が進められているものの，LLM によって出力された回答を評価する指標については課題が残る．

現在，質問応答タスクにおいては，参照回答と出力回答の一致度を測る指標として，ROUGE スコア [6, 8, 7] や BLEU スコア [10] が広く用いられている．一方でこれらの手法では捉えられない観点が存在し，文献 [1, 9] では文字や単語だけではなく意味的な一致度の評価が必要だと示されている．

そのような背景に基づき，意味的な一致度を重視した評価指標としては，BERT スコア [16] や COMET スコア [11]，および，文献 [2, 5] が提案されている．さらに文献 [15] では参照回答を必要とせずに出力回答を評価できる指標を提案しており，文献 [12] では LLM の分散表現から品質を表した射影を抽出することで回答を評価している．

しかしこれらの評価指標では出力回答に含まれる構造的な重要度のほか，情報が不足している場合や情報が不適切である場合について評価することができない．そこで文献 [14] では参照回答と出力回答をそれぞれ「趣旨」「補足」と構造的に分解し，さらにそれらを比較し「不適」「不足」である箇所を数値化して評価することで，新たな観点からの評価を実施したが，これらは人手による作業で評価されていたため，大規模な評価には適していなかった．そこで本研究においては，参照文の構造的な分解と「不適」「不足」の評価を LLM で自動化することによって，より効率的な評価を可能とした．

3　回答中の趣旨・補足情報

質問応答タスクにおいては，正解である参照回答により近い出力回答を得ることを目標としている．

参照回答の趣旨・補足の同定指示プロンプト
あなたはプロの会計士です。
これから会計情報に関する質問１つとそれに対する参照回答（正解データ）を与えます。
与えられた参照回答について質問に対する答えとして重要な「趣旨」とそうでない「補足」情報を区別してください。

手順
1．質問文を見て何を回答するべきかを把握する
2．参照回答の「趣旨」「補足」情報で参照回答を分割する。分割単位は文ではなく、意味単位（意味的に独立した情報）で計測する。
以下の条件を再度見直し、問題ないことを確認する

条件
・区切りは文単位や改行では無いことに注意（意味単位で区切る）
・全ての参照回答は「趣旨」と「補足」情報のどちらかに必ず属するものとする。
・「趣旨」とは質問に対する答えとして、参照回答や出力回答の中で特に重要であると判断できる情報
・「補足」とは質問に対する答えとして、参照回答や出力回答の中で「趣旨」を補足している情報や具体例、重要性が低い情報

出力回答の不適・不足の検出指示プロンプト
あなたはプロの会計士です。
これから会計情報に関する質問１つとそれに対する参照回答（正解データ）が１つ、LLMの出力回答を1つ与えます。
出力回答に対して参照回答との一致度を測ってください。
一致度の計測には参照回答や出力回答中の「趣旨」「補足」情報で区別して行います。詳しい条件や手順は以下に示します。

手順
1. 出力回答と参照回答を比較し、「不適」と「不足」の情報を検出する。言い換え等、大まかな意味があっていれば一致しているとみなします。
2. 参照回答の「趣旨」「補足」のどちらに対応しているか判別する。「不適」「不足」それぞれに対しこれを行うため、計４通りの組み合わせがある。

条件
・「趣旨_不適」「趣旨_不足」「補足_不適」「補足_不足」に同じ要素は入れないでください
・「趣旨」とは質問に対する答えとして、参照回答や出力回答の中で特に重要であると判断できる情報
・「補足」とは質問に対する直接の答えではなく、参照回答や出力回答の中で「趣旨」を補足している情報や具体例、重要性が低い情報
・「不適」とは参照回答には書かれていないが、出力回答で書かれている情報
・「不足」とは参照回答には書かれているが、出力回答では言及されていない情報

図 1　LLM による不適・不足検出の指示プロンプト

しかし参照回答には重要度の高い情報だけではなく重要度の低い情報が存在していると考えられ，評価に際しその違いを考慮した指標が必要となる．そこで文献 [14] においては，参照回答を主張の中心となる事柄であり重要度の高い「趣旨」と，それを補う情報や具体例などの相対的に重要度の低い「補足」の二種類に分類して評価する指標が提案された．本論文においても同じ定義に則って参照回答を「趣旨」と「補足」に分類する．

4　回答中の不適・不足

出力回答は参照回答に対して情報の過不足が発生することがある．文献 [14] においては，参照回答よりも情報が多い場合を「不適」，参照回答よりも情報が少ない場合を「不足」とし，これを前節の「趣旨」「補足」の分類と組み合わせることで，以下の四種類の特性を定義した．

- 「趣旨」が「不適」
- 「趣旨」が「不足」
- 「補足」が「不適」
- 「補足」が「不足」

なお，「不適」とは参照回答には書かれていない真偽不明の情報がつけ足されている状態であり，その情報が誤りである場合と明確な誤りではないが冗長である場合のふたつの可能性がある．3 節と同様に，本論文においても同じ定義を用いて出力回答中の「不適」「不足」を分類する．

5　回答中の不適・不足の評価指標

文献 [14] ではさらに，出力回答中の「不適」「不足」の評価指標を定義した．まず参照回答中の質問の答えとして重要な情報を「趣旨」，「趣旨」の補足や具体例，その他の重要度の低い情報を「補足」と分類し，それらの情報を意味単位で区切り個数を計測した．

次に参照回答の分解された「趣旨」「補足」ごとに，評価対象となる回答が「不適」であるか「不足」であるか，それぞれ二値で判定した．そして「不適」と判断された個数を「趣旨」の個数で割ることで，「趣旨」における「不適」の評価値とした．同様の数値を前節で示した四種類の特性において計測し，その評価値に応じて回答を分類した．

以上の手順を踏まえ，各評価値が 0 に近い方がより重要な情報を的確に回答できたことを評価する．

6　LLM による不適・不足の自動検出

本節では 4 節で示した「趣旨」「補足」情報毎の「不適」「不足」を LLM によって検出する手法について述べる．

6.1　参照回答中の趣旨・補足の同定

LLM による「不適」「不足」の自動検出の前段階の手順として，3 節で示した参照回答中の「趣旨」と「補足」情報の同定を行う．同定にあたって，LLM に参照回答を入力し，参照回答中の文字列を意味単

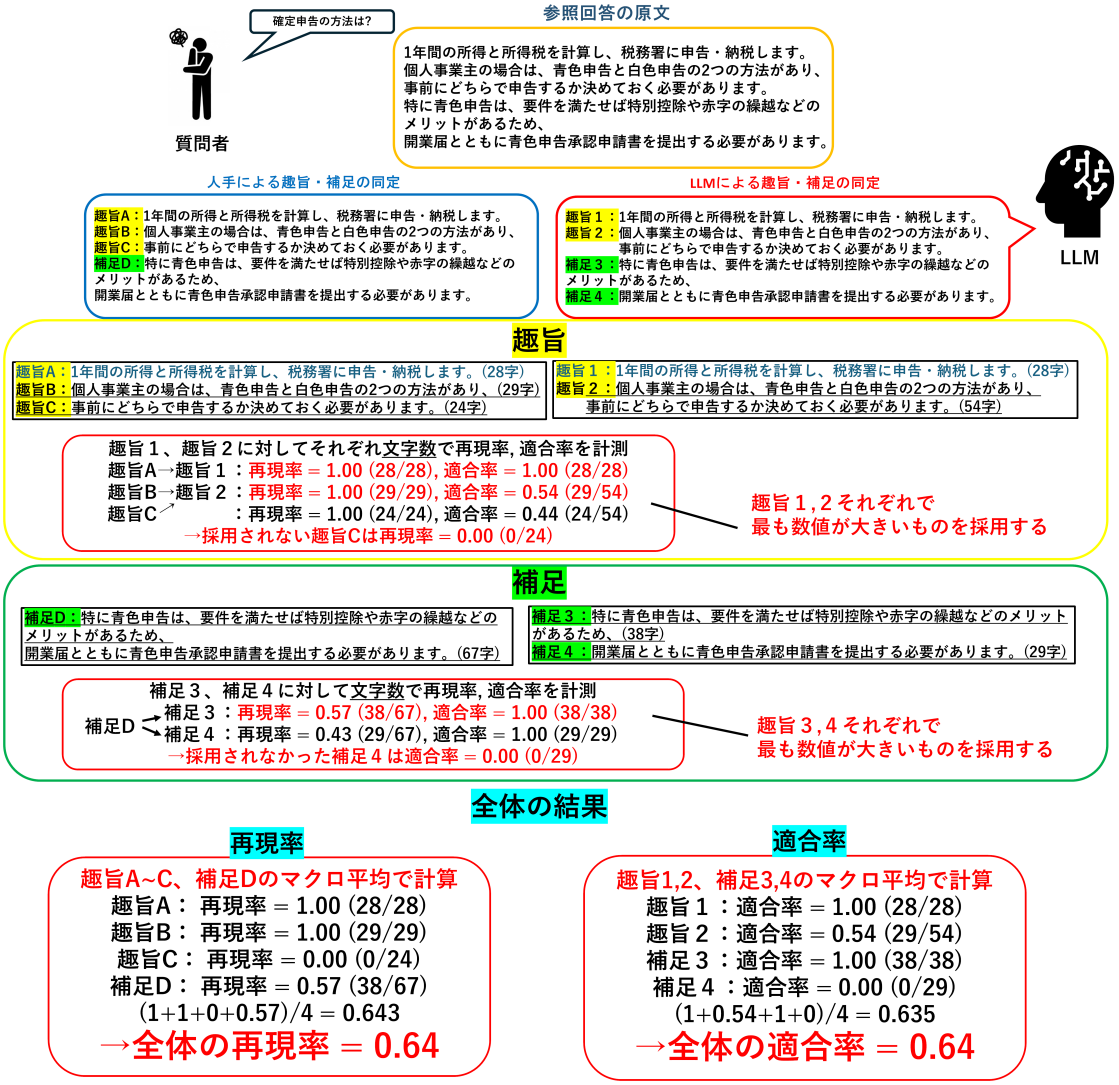

図 2　参照回答の趣旨・補足情報の同定評価手順

位で分解し，各単位ごとに「趣旨」「補足」を区別するよう指示を与える．この手順で得た参照回答中の「趣旨」「補足」情報の同定結果は 6.2 節で行う「不適」「不足」の自動検出に使用する．

6.2　趣旨・不足情報の不適・不足の検出

6.1 節で行った参照回答の「趣旨」「補足」情報の同定結果と，出力回答の「趣旨」「補足」情報ごとに「不適」「不足」の箇所を LLM により検出する．この際 5 節で示したように「不適」には誤情報と冗長な情報の 2 通りがあるが，本論文はその区別は実施しない．LLM は「不適」「不足」箇所を抽出し，さらにその箇所が「趣旨」「補足」のいずれであるか 6.1 節の情報から判断するため，抽出結果は 4 節で紹介した 4 種類の特性のいずれかに分類される．また前節で行う参照回答中の「趣旨」「補足」の同定から本節の「不適」「不足」検出までを行うプロンプトの全文を図 1 に示す．

表 1　参照回答・出力回答の比較結果における，人手・LLM による検出結果が存在しない組み合わせの数

組み合わせ数	0	1	2	3	4	計
趣旨・補足の同定を人手で行った場合	18	8	13	7	4	50
趣旨・補足の同定を LLM で行った場合	7	12	17	9	5	50

7　評価

7.1　評価手順

LLM による参照回答の「趣旨」「補足」情報の同定や，「不適」「不足」の検出の精度について，人手検出した文と LLM が検出した文を比較し，差異がある箇所の文字数を元に計算する．

表 2　参照回答・出力回答の比較結果における不適・不足自動検出の評価結果

	不適						不足					
	趣旨			補足			趣旨			補足		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
趣旨・補足の同定を人手で行った場合	0.26	0.13	0.18	0.48	0.49	0.49	0.44	0.30	0.36	0.53	0.41	0.46
趣旨・補足の同定を LLM で行った場合	0.24	0.15	0.18	0.53	0.58	0.56	0.33	0.50	0.40	0.48	0.47	0.47

詳細な手順について図 2 で説明する．ここでは参照回答の原文に対して人手で「趣旨」「補足」情報を区別した結果を左，LLM で同様の作業を行った結果を右に示している．まず趣旨について，人手では 3 件得られた一方、LLM では 2 件のみ抽出している．このとき趣旨の件数が一対一になるような対応を取り，その上で文字単位の再現率と適合率を計算する．たとえば趣旨 A と趣旨 1 は文字数まで完全に一致しており，再現率と適合率は 1.00 になる．一方，趣旨 B と趣旨 C は共に趣旨 2 に対応しているため，両者の再現率・適合率を計算し，より数値が大きい趣旨 B を趣旨 2 に対応するものとして採用する．この時，対応先のない趣旨 C は再現率が 0，適合率は分母が 0 となるため計測不可となる．補足や不適・不足の検出性能評価の際も同様の計算を行う．

以上の評価手順に従うことで，{ 不適, 不足 } × { 趣旨, 補足 } = 4 通りの組み合わせの各組のうち，人手による検出が無い組は出力回答と参照回答が一致しており，LLM による検出の必要がないため再現率が計測不可となる．一方，4 通りの組み合わせの内，LLM による検出が得られなかった組については，適合率が計測不可となる．

さらに，人手による検出結果も LLM による検出結果も存在しない場合も考えられる．これは「不適」「不足」の箇所がなく適切な出力回答が得られつつ，さらに人手・LLM いずれも「不適」「不足」と誤検出しなかった場合で，再現率・適合率は共に計測不可となるため，そのような回答は別で集計を行う．また，以上の「不適」「不足」検出に関するより詳細な説明として，付録の図 3 に示す．

7.2　評価結果

まず 4 節で示した { 不適, 不足 } × { 趣旨, 補足 } = 4 通りの組み合わせの中で，前節で述べた人手による検出結果も LLM による検出結果も存在しない組について集計した表を表 1 に示す．次に「不適」「不足」自動検出を人手・LLM で実施した際の評価結果を表 2 に示す．このとき，表 1 で集計した人手・LLM ともに検出結果が存在しなかった組は，再現率・適合率の計測ができなくなるため除外する．また再現率・適合率はマクロ平均で計算し，F 値はマクロ平均後の再現率・適合率から導出した．

表 1 から，人手評価で「不適」「不足」が存在しない回答については LLM でも同様に「不適」「不足」が存在しないと正しく判断できていることが分かる．さらに，表 2 の結果を見ると，「不適」「不足」が存在した場合，その「不適」「不足」部分についておよそ半分ほどが正しく検出できている．また，「不適・趣旨」は他よりも精度が大幅に小さくなっているが，「不適」については LLM が過剰に付け足す部分であるため，あらかじめ「趣旨」「補足」を同定することが出来ず，「趣旨」のような大きな情報の「不適」を検出することは困難であることを示している．

しかし本論文のプロンプトは人手で行う評価作業を手順化しただけの単純なものであり，Few-Shot Learning の導入をはじめ，検出困難な箇所へ向けて改良する余地が十分に残されていると言える．

8　おわりに

本論文では質問応答タスクにおいて，回答中の「趣旨」「補足」情報まで考慮した評価を行い，出力回答が参照回答と比較して過剰に書かれている箇所や欠けている箇所をそれぞれ「不適」「不足」と定義し，出力回答を詳細に分類した．

さらに，前述の検出手順を人手ではなく LLM を用いて自動的に実施する手法を提案し，LLM による検出結果を文献 [14] で実施された人手評価による検出結果と比較することで，本論文の自動評価に人手評価との大きな差異が観察されなかったことを示し，文献 [14] で提案された評価を大規模に実行可能とした．

参考文献

[1] M. Akter, N. Bansal, and S. K. Karmaker. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In **Findings of ACL**, pp. 1547–1560, 2022.

[2] M. Gao, J. Ruan, R. Sun, X. Yin, S. Yang, and X. Wan. Human-like summarization evaluation with ChatGPT. **arXiv preprint**, Vol. arXiv:2304.02554, , 2023.

[3] 飯田頌平, 古俣槙山, 三田寺聖, 長谷川遼, 宇津呂武仁, 林友超, 宍戸里絵. RAG に基づく会計分野の質問応答. 第 39 回人工知能学会全国大会論文集, 2025.

[4] 飯田頌平, 古俣槙山, 三田寺聖, 長谷川遼, 宇津呂武仁, 林友超, 宍戸里絵. 会計ドメインにおける質問応答のための LLM を用いた解説ページ順位付け. 言語処理学会第 31 回年次大会論文集, pp. 2757–2762, 2025.

[5] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In **Proceedings of EMNLP**, pp. 4334–4353, November 2024.

[6] C. Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. 2003.

[7] C. Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. 2004.

[8] C.Y. Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[9] F. M. Molfese, L. Moroni, L. Gioffré, A. Scirè, S. Conia, and R. Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In **Findings of ACL**.

[10] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of ACL**, July 2002.

[11] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of EMNLP**, pp. 2685–2702, 2020.

[12] S. Sheng, Y. Xu, T. Zhang, Z. Shen, L. Fu, Jiaxin Ding, L. Zhou, X. Gan, X. Wang, and C. Zhou. RepEval: Effective text evaluation with LLM representation. In **Proceedings of EMNLP**, November 2024.

[13] 高橋空大, 土田陸斗, 三田寺聖, 謝宇程, 長谷川遼, 宇津呂武仁, 飯田頌平, 林友超, 宍戸理恵. LLM による会計ドメイン質問応答における RAG の有効性の評価. 第 18 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2026.

[14] 土田陸斗, 飯田頌平, 高橋空太, 三田寺聖, 長谷川遼, 謝宇程, 宇津呂武仁, 林友超, 宍戸里絵. QA タスク回答中の趣旨・補足に対する不適・不足の分析. 言語処理学会第 31 回年次大会発表論文集, 2026.

[15] A. Wang, K. Cho, and M. Lewis. Asking and answering questions to evaluate the factual consistency of summaries. 2020.

[16] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTscore: Evaluating text generation with BERT. In **ICLR Workshops**, 2020. Preprint available at arXiv:1904.09675.

A　出力回答の不適・不足自動検出の評価手順

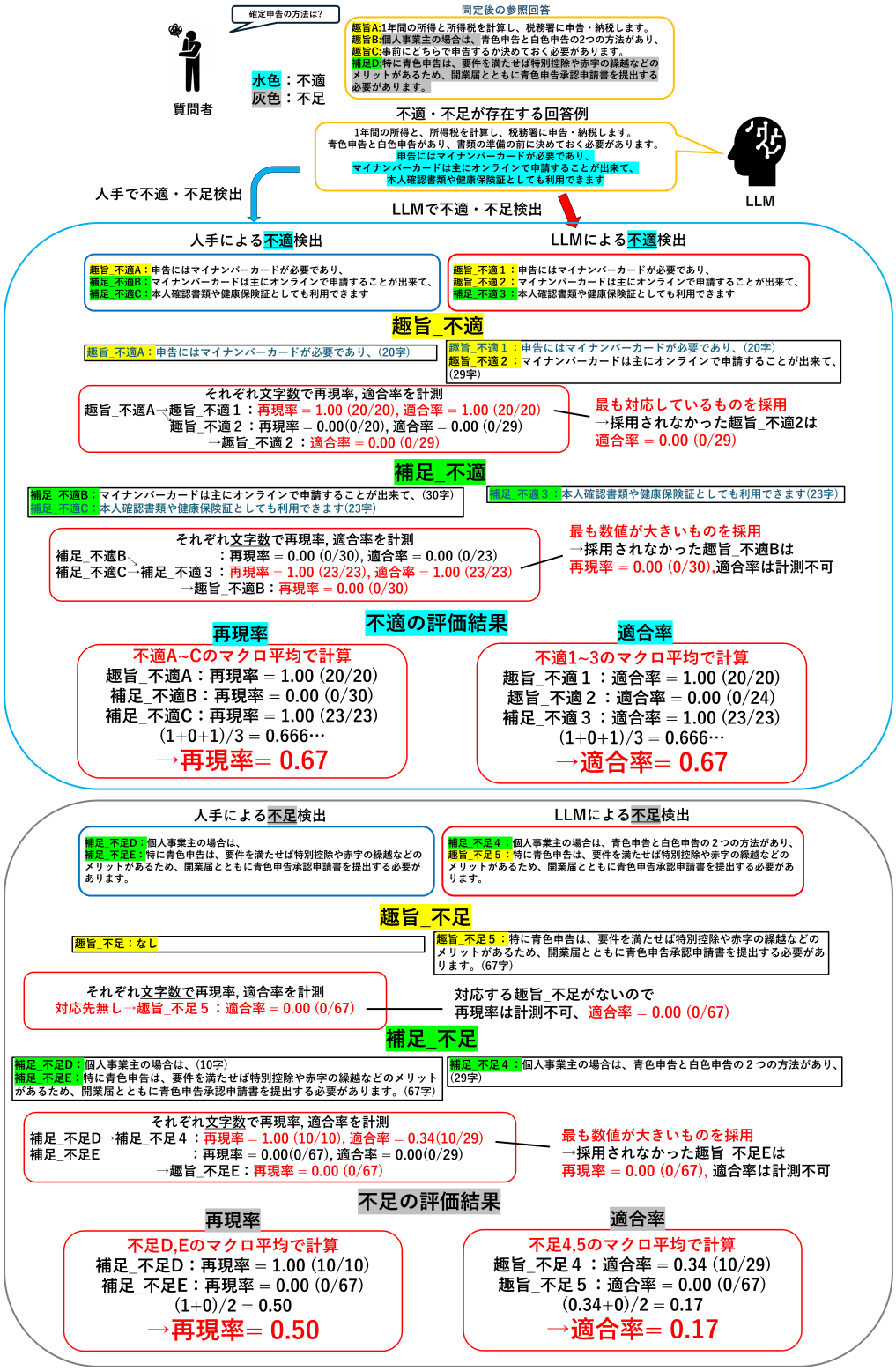

図 3　出力回答の不適・不足自動検出の評価手順