

採点基準の分解とデータ拡張による 要約ライティング自動採点

鈴木颯仁¹ 山田寛章¹ 徳永健伸¹
石井雄隆² 藤田祐美³ 澤木泰代³
¹ 東京科学大学 ² 千葉大学
³ 早稲田大学

suzuki.h.2e97@m.isct.ac.jp yamada@comp.isct.ac.jp take@c.titech.ac.jp
yishii@chiba-u.jp yumifuji.edu@fuji.waseda.jp ysawaki@waseda.jp

概要

本研究では、教育現場における要約ライティングの自動採点を目的として、大規模言語モデル (LLM) を用いた三段階の自動採点手法を提案する。本手法は、採点基準の分解、サンプルの拡張、拡張サンプルを参照した採点から構成される。二種類の学習者要約データセットを用いた実験の結果、fine-tuning を行わずとも、既存手法より人手採点に近い性能を達成できることを示した。

1 はじめに

第二言語習得のための英語教育において、要約ライティングは重要な教材である。文章要約の能力は学術的活動においても必要性が高く [1]、実践的な課題として授業内で利用される。また、第二言語作文課題では、学習者が課題を提出した直後に採点結果を提供する「即時フィードバック」が、流暢さや文構造の豊かさなどを高める効果があるとする研究 [2] も報告されている。しかし、学習者の要約を採点し、適切なフィードバックを与えることは多大な負担であることから、人間の教師が即時にフィードバックを与えることは難しい。

近年、LLM を活用した自動採点が多数提案されているが [3]、これらをそのまま学生の要約採点に適用することは、以下の二つの理由から難しい。

- 多くの自動採点手法は、主に語彙や文法といった文章の言語的側面に焦点が当てられるのに対して [4, 5]、要約採点においては、要約対象のテキストの内容に即した内容的側面を重視した採点の方が適している。
- 一部の自動採点手法では大量の学習データを必

要とする機械学習アプローチが採用されているが [6]、授業で用いられる教材は一度に利用する人数が限られているうえ、授業の進行や学習目的に応じて題材が柔軟に変更されることが多いため、特定の教材に対して十分な量の学習データを継続的に収集することは困難である。

これらの課題に対処するために、本研究では採点基準を分解するとともに、few-shot サンプルを拡張して文脈内学習の質を高める、LLM を用いた新しい要約自動採点手法を提案する。

2 関連研究

人手生成文の採点を扱った研究として、Li ら [7] は質問応答を対象に、Yancey ら [4] はエッセイを対象として、LLM が人手による採点に代替し得る可能性を示している。

一方で、自然言語生成の採点を扱った研究として、Liu ら [8] は、LLM を用いて採点基準をより詳細な採点ステップへと展開する自動 Chain-of-Thought を導入し、採点手順を明示的に与える G-Eval を提案した。その検証実験の結果、G-Eval は BLEU や ROUGE といった従来の自動採点指標と比較して、人手採点との整合性が高いことを示した。この手法を拡張した Chiang ら [9] は、最終的な点数に加えて採点の根拠を出力させることで、採点の安定性および人手採点との一致度がさらに向上することを報告している。また、採点性能向上のために採点基準を分解するアプローチも提案されている。Liu ら [10]、Wei ら [11] は、採点基準を複数の側面に分解し、各側面を個別に採点した上で統合することで、高精度な採点を実現している。これらの手法では、採点結果を統合するために学習されたモジュールを必要と

する点が特徴である。

採点タスクを直接の対象とした研究ではないものの、Suら [12] は、過去にモデル自身が予測したサンプルを「デモンストレーション」として再利用する手法を提案し、外部の追加データをほとんど用いずに、few-shot 学習と同等、あるいはそれ以上の性能を達成できることを示し、LLM の推論能力を効率的に引き出すための知見を提供している。

3 提案手法

本研究では、図 1 に示すように、LLM を用いた三段階からなる要約自動採点手法を提案する。本手法は、採点基準を分解して質問を生成することで採点の曖昧さを低減し、さらに、人手で作成することが困難な、詳細な採点過程を付与した拡張サンプルを LLM により自動構築する。このサンプルを参照することで、教室環境において実用的かつ一貫性のある要約採点を実現することを目的とする。

本手法の設計に先立ち、GPT-4.1 を用いて複数の予備実験を実施し、要約採点における LLM の挙動を検証した。その結果、(i) 採点基準の分解を具体的な条件を与えずに指示した場合、採点基準に明記されていない観点まで推測的に含めてしまう傾向がある、(ii) 採点基準以外の具体的な条件や指示を与えずに採点を行わせた場合、評価・改善提案・助言などが混在した出力が生成され、どの観点による採点か不明瞭になる傾向がある、(iii) 学習者の要約に対して採点基準に沿った分析と点数を生成させた際、LLM が推定した点数と教師が実際に付けた点数との間に乖離が生じ、分析結果の正確性が低下することがある、の三点が確認された。これらの結果を踏まえ、本研究では、各段階において LLM の出力に制約や補助を加える設計を採用する。

第一段階では、要約対象のテキストと採点基準を入力とし、採点基準を分解した採点要素と、それぞれの要素を質問形式に変換した質問を生成することで、採点基準の分解を行う。予備実験から得られた知見 (i) より、本研究では採点基準に明示された情報のみを用いて分解する制約を設けている。さらに知見 (ii) より、各採点要素を具体的かつ限定的な質問へ変換することで、LLM が個々の採点要素に対して焦点を絞った判断を行えるように設計している。

第二段階では、要約対象テキスト、採点基準、第一段階で生成した質問に加え、サンプルデータに含まれる学習者の要約と人手による点数を入力として

用いる。LLM には、人手による点数を保持したまま、各質問への回答およびそれらを踏まえた全体の分析を生成させ、これらをサンプルデータに追加することでデータ拡張を行う。予備実験により得られた知見 (iii) より、質問への回答および全体の分析を生成する際に、人手による点数を明示し、正確性と信頼性を確保する。

第三段階では、要約対象のテキスト、採点基準、第一段階で生成した質問、および採点対象となる学習者の要約に加え、第二段階で作成した拡張サンプルデータを採点例として入力し、採点を行う。こうした参照プロセスにより、各質問に対する判断基準が明示的に共有されるため、採点要素間およびサンプル間での判断の揺らぎを抑制できる。その結果、採点基準に忠実で一貫性のある要約採点が可能となる。

以上の三段階からなる手法により、本研究は、fine-tuning を必要とせず、限られたデータ環境においても高品質な要約自動採点を実現する枠組みを提案する。

4 実験設定

4.1 データセット

本研究では、飲料水問題とその解決策を題材とした Cycloclean (391 語) と仮眠の効果と利点を題材とした Napping & Learning (N & L) (415 語) の 2 種類の英語文章の要約 [13, 14, 15, 16] を扱う。各データセットの統計情報を表 1 に示す。採点の客観性と信頼性を担保するため、二名の英語教師による採点后、協議を経て最終的な合意点数協議が付与された 60 件をそれぞれ評価データとして使用し、残りを学習データ、開発データとして分割して用いる。

採点基準 それぞれの要約対象のテキストに対して、要約の内容の適切性を多角的に採点するための 13 個の採点基準が定義されている。

要約 要約対象のテキストを読み、そのポイントとなる情報を 80 語程度で要約するという授業内課題の一環で収集された。

4.2 実験条件

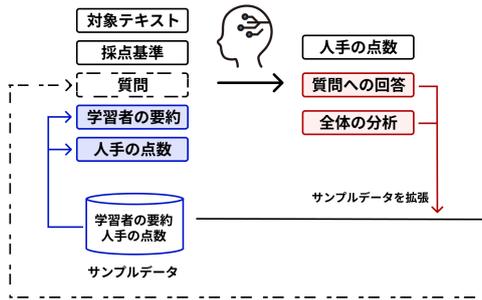
本研究では以下の条件で実験を行う。

入力 要約対象のテキストと 13 個の採点基準、評価データの要約を入力として与える。

第一段階 採点基準の分解と質問の生成



第二段階 サンプルデータの拡張



第三段階 拡張サンプルを参照した採点

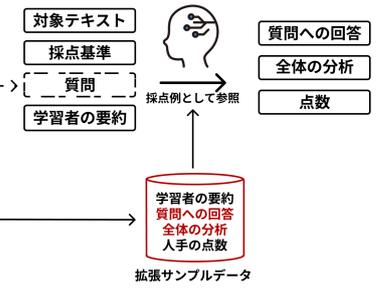


図 1 提案手法の概要図

項目	Cycloclean	N & L
総データ数	132 件	150 件
要約語数	約 80 語	約 80 語
大学 A 作成 (年別)	2020 年: 40 件 2023 年: 25 件 2024 年: 35 件	2023 年: 58 件 2024 年: 30 件
大学 B 作成 (年別)	2021 年: 32 件	2021 年: 32 件 2022 年: 30 件
二者採点	60	60
単独採点	72	90
評価データ	60	60
開発データ	30	30
学習データ	42	60

表 1 Cycloclean および N & L データセットの統計情報. 総データ数, 要約語数, 作成大学・年別内訳, 採点形態およびデータ分割を示す.

出力 0, 1, 2 の三段階の点数と, その他根拠等を出力として生成する.

使用モデル 表 2 に示す, 四種類の商用モデルと二種類のオープンモデルを用いる.

ベースライン 提案手法の有効性を示すために, 以下の五つのベースラインを設定する.

- **Zero-shot**: モデルに対して採点例となるサンプルを一切与えずに採点を行う.
- **Few-shot**: 採点対象の要約と, それに対して人手で付与された点数をサンプルとしてモデルに与える. 使用するサンプルは, いずれの手法においても共通して, 学習データの中から各採点基準において点数が 0, 1, 2 となるサンプルをランダムに一件ずつ抽出したものである.
- **G-Eval**[8, 9]: 採点基準を LLM 自身が展開して具体的な採点ステップを生成し, そのステップに従って点数を生成する.
- **G-Eval Few-shot**: G-Eval において点数生成時に, 採点対象の要約とそれに対して人手により

区分	モデル	設定・パラメータ
商用	GPT-4.1 GPT-5 [17] Claude S 4.5 [†] [18] Claude S 4.5 [†] ET	temp= 0, seed= 20598126 reasoning effort: medium - extended thinking: enabled
オープン	gpt-oss-20B [19] gpt-oss-120B [19]	reasoning effort: medium reasoning effort: medium

表 2 使用モデル一覧. [†] は Claude Sonnet 4.5 を指す.

付与された点数をサンプルとして与える.

- **提案手法 (拡張なし)**: 提案手法の構成要素のうち, サンプル拡張を行わず, 採点対象となる要約とそれに対して人手により付与された点数のみをサンプルとして与える.

評価方法 13 個の採点基準ごとに, Micro-F1 スコアを計算し, それらの平均値を全体評価する. さらに, 手法間の全体評価の差に対して, 有意水準 0.05 の両側検定による並び替え検定を行う.

5 結果

実験結果を表 3 に示す. 商用モデルによる実験では, 提案手法の評価値は一貫して他のベースライン手法を上回っており, 安定した性能向上が確認された. このうち, GPT-5 を用いた Cycloclean での実験および Claude Sonnet 4.5 ET を用いた実験を除き, 全ての実験で, 提案手法がベースライン手法と比較して統計的に優れた性能を示した.

一方, オープンモデルを用いた実験では, モデルサイズ間の性能差が顕著に観察された. gpt-oss-20B では, Zero-shot および Few-shot が提案手法を上回る結果になったのに対し, より大規模な gpt-oss-120B では, 提案手法がベースライン手法に対して統計的に優れた性能を示した.

この結果は, 提案手法が一定水準以上の推論能力を備えたモデルを前提として有効に機能する可能性

モデル	手法	Cycloclean F1	N & L F1
GPT-4.1	Zero-shot	0.613*	0.563*
	Few-shot	0.619*	0.555*
	G-Eval	0.622*	0.567*
	G-Eval Few-shot	0.630*	0.561*
	提案手法 (拡張なし) 提案手法	0.658* 0.701	0.577* 0.603
GPT-5	Zero-shot	0.709*	0.600*
	Few-shot	0.727	0.604*
	G-Eval	0.703*	0.609*
	G-Eval Few-shot	0.713	0.577*
	提案手法 (拡張なし) 提案手法	0.724 0.746	0.625* 0.664
Claude Sonnet 4.5	Zero-shot	0.601*	0.526*
	Few-shot	0.619*	0.615*
	G-Eval	0.700*	0.614*
	G-Eval Few-shot	0.688*	0.635*
	提案手法 (拡張なし) 提案手法	0.713* 0.750	0.642 0.676
Claude Sonnet 4.5 ET	Zero-shot	0.726	0.614*
	Few-shot	0.726	0.619*
	G-Eval	0.718	0.669
	G-Eval Few-shot	0.721	0.662*
	提案手法 (拡張なし) 提案手法	0.727 0.737	0.660* 0.694
gpt-oss- 20B	Zero-shot	0.625	0.586
	Few-shot	0.639	0.589
	G-Eval	0.598	0.567
	G-Eval Few-shot	0.586	0.571
	提案手法 (拡張なし) 提案手法	0.584 0.567	0.563 0.541
gpt-oss- 120B	Zero-shot	0.618*	0.598*
	Few-shot	0.635*	0.598*
	G-Eval	0.689*	0.610*
	G-Eval Few-shot	0.671*	0.579*
	提案手法 (拡張なし) 提案手法	0.700 0.726	0.619* 0.655

表3 各モデルにおける各手法の評価値の比較。*は提案手法との間に有意差があることを示す。太字で示された数値は、各モデルごとの最も高い評価値を示している。

を示唆している。ここでいう推論能力とは、入力されたタスクの意図を正確に理解し、一貫性と信頼性のある出力を生成する能力を含む総合的な性能を意味する。

本手法では、モデル自身が生成した拡張サンプルを後段の推論で参照するため、生成される拡張サンプルの品質が採点性能に直接的な影響を及ぼす。推論性能の低いモデルでは、生成される回答に不正確さが増す傾向があり、本手法の第三段階における参照される拡張サンプル自体の品質が低下する。その

結果、LLM が不適切なサンプルを参照して採点を行い、かえって性能が低下したことが考えられる。先行研究 [12] においても、few-shot 学習において提示されるサンプルの品質が性能に大きく影響することが報告されている。

次に、提案手法 (拡張なし) に着目することで、本手法を構成する各要素の有効性を分析する。提案手法 (拡張なし) を Few-shot, G-Eval Few-shot と比較した結果、Claude Sonnet 4.5 を用いた N & L での実験を除くすべての条件において提案手法 (拡張なし) がより高い評価値を示した。この結果から、第一段階における採点基準の分解および質問化が、採点性能向上に寄与していることが確認された。さらに、提案手法 (拡張なし) と提案手法を比較すると、全体として提案手法の方が高い性能を示す傾向が確認された。このことから、第二段階で行うサンプル拡張が、採点性能向上に一定の効果を持つと考えられる。一方で、gpt-oss-20B を用いた場合には評価値が大きく低下する結果となった。これは、前述のとおり、用いたモデルの性能不足により低品質な拡張サンプルが生成され、その結果、採点性能が悪化したためであると考えられる。

6 結論

本研究では、採点基準の分解と質問化、サンプル拡張、および拡張サンプルを参照した採点からなる三段階の要約の自動採点手法を提案し、その有効性を示した。特に、fine-tuning を必要とせず、few-shot 学習のみで人手採点に近い性能を実現できる点は、教育現場における実用性の観点から重要な貢献である。

今後の課題としては、まず質問への回答や全体分析の内容を詳細に分析し、人手採点との差異が生じる要因をより精緻に考察することが挙げられる。また、事例のサンプリングにおいて、単純なランダムサンプリングではなく、質問への回答や分析の類似度に基づく選択手法を導入することで、更なる性能向上が期待される。加えて、本手法を実際の教育支援システムへ組み込み、教室環境での有効性を検証することも重要な課題である。今後は、要約採点以外の分野への応用可能性についても検討していく予定である。

謝辞

本研究は JSPS 科研費 JP20H01292, JP24K00095 の助成を受けたものです。

参考文献

- [1] Ilona Leki and Joan Carson. "Completely Different Worlds": EAP and the Writing Experiences of ESL Students in University Courses. *TESOL Quarterly*, Vol. 31, No. 1, pp. 39–69, 1997.
- [2] Grant Eckstein, Maureen Sims, and Lisa Rohm. Dynamic written corrective feedback among graduate students: The effects of feedback timing. *TESL Canada Journal*, Vol. 37, No. 2, p. 78–102, Dec. 2020.
- [3] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.
- [4] Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 576–584, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Changrong Xia, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint arXiv:2401.06431*, 2024.
- [6] Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13034–13054, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [7] Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. Distilling ChatGPT for explainable automated student answer assessment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6007–6026, Singapore, December 2023. Association for Computational Linguistics.
- [8] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.
- [9] Cheng-Han Chiang and Hung yi Lee. A Closer Look into Using Large Language Models for Automatic Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8928–8942, Singapore, 2023. Association for Computational Linguistics.
- [10] Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. HD-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7641–7660, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [11] Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. Rocketeval: Efficient automated LLM evaluation via grading checklist. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [12] Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Yan Bowen, and Min Zhang. Demonstration augmentation for zero-shot in-context learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14232–14244, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [13] Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki, and Mika Ishizuka. Automating idea unit segmentation and alignment for assessing reading comprehension via summary protocol analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4663–4673, Marseille, France, June 2022. European Language Resources Association.
- [14] Yasuyo Sawaki, Yutaka Ishii, Hiroaki Yamada, and Takenobu Tokunaga. Developing and validating an online module for formative assessment of summary writing with automated content feedback for EFL academic writing instruction. *Language Testing in Asia*, Vol. 14, No. 50, 2024.
- [15] Yasuyo Sawaki, Yutaka Ishii, and Yoko Oi. The efficacy of automated feedback on summary content in an online formative assessment module for undergraduate L2 writing instruction in Japan. *Studies in Language Assessment*, Vol. 13, No. 2, pp. 38–72, 2024.
- [16] Yasuyo Sawaki, Yutaka Ishii, Hiroaki Yamada, and Takenobu Tokunaga. Examining the consistency of instructor versus large language model ratings on summary content: Toward checklist-based feedback provision with second language writers. *Language Testing*, Vol. 42, No. 4, pp. 447–475, 2025.
- [17] OpenAI. GPT-5 System Card. <https://cdn.openai.com/gpt-5-system-card.pdf>, August 2025. Accessed: 2025-12-26.
- [18] Anthropic. System Card: Claude Sonnet 4.5. <https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf>, September 2025. Accessed: 2025-12-26.
- [19] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, et al. gpt-oss-120b & gpt-oss-20b model card, 2025.

A 提案手法における入出力の例

A.1 第一段階

入力：採点基準

Is “water for drinking, cooking, or washing” in Paragraph 1 summarized into a phrase such as “water for living”, “purified water”, “drinking water”, “safe water”, and “clean water” in the summary?

出力：採点要素

- C1: The summary includes the concept of water for multiple purposes (drinking, cooking, washing) as mentioned in Paragraph 1
- C2: The summary condenses or generalizes the specific list "drinking, cooking, or washing" rather than listing each use individually
- C3: The generalization uses appropriate terminology, such as the examples provided: "water for living", "purified water", "drinking water", "safe water", "clean water", or equivalent expressions

出力：質問

- Q1: Does the summary include reference to the water needs mentioned in Paragraph 1?
- Q2: Does the summary avoid listing "drinking, cooking, or washing" as separate items and instead use a more generalized expression?
- Q3: If the summary generalizes the phrase, does it use an appropriate term such as "water for living", "purified water", "drinking water", "safe water", "clean water", or a similar suitable expression?

A.2 第二段階

入力：学習者の要約

Running out of clean water is a serious problem; it leads to the death of many children according to a research. For the sake of solving this, a product named Cycloclean which enables people to change dirty water into clean one was made. It can make a large amount of water in any place where unsafe water and an individual are. However, the weaknesses are its number and price. Because of this, availability of it is confined, but now that companies deal with this, Cycloclean will prevail all over the world in the near future.

入力：人手の点数

Score: 2

出力：質問への回答

- A1. Yes, the summary includes reference to the water needs mentioned in Paragraph 1. The opening sentence states "Running out of clean water is a serious problem," which directly addresses the water needs described in the source text's first paragraph about "safe water for drinking, cooking, or washing."
- A2. Yes, the summary successfully avoids listing "drinking, cooking, or washing" as separate items. Instead of enumerating these specific uses, the summary employs the generalized expression "clean water" throughout, including phrases such as "Running out of clean water" and "change dirty water into clean one."
- A3. Yes, the summary uses an appropriate generalized term. It employs "clean water" multiple times (e.g., "Running out of clean water," "change dirty water into clean one"), which is explicitly listed as one of the suitable expressions in the evaluation criteria. This term effectively captures the essence of the original phrase without unnecessarily detailing the specific uses.

出力：全体の分析

The summary fully meets the evaluation criteria by successfully condensing the specific phrase "water for drinking, cooking, or washing" from Paragraph 1 into the generalized term "clean water." The summary demonstrates appropriate summarization skills by avoiding unnecessary detail while maintaining the core meaning. The term "clean water" appears naturally integrated into the summary's opening statement and is consistently used throughout, showing that the student understood the need to generalize rather than enumerate specific water uses. This approach aligns perfectly with the criteria's expectation for using terms such as "clean water," making the 2-point score appropriate.

A.3 第三段階

第二段階と同様の入出力形式であるため省略する。