英語長文穴埋め問題のための問題カテゴリ推定と空欄箇所選定手法の改善

板橋 康知 [1,2]　松林 優一郎 [1,2]
[1] 東北大学 [2] 理化学研究所
itabashi.yasutomo.q7@dc.tohoku.ac.jp　y.m@tohoku.ac.jp

概要

本研究では、英語長文穴埋め問題で教育目的上重要となる長期推論型の問題を自動生成することを目的として、その基盤となる作成された問題のタイプを自動分類する評価手法を確立する。また、提案する評価手法で教師の作成した問題と既存の生成手法で生成された問題タイプの分布を詳細に分析し、既存手法の問題点を明らかにする。さらに、この結果をもとに、既存の問題生成手法のさらなる改良を試み、その結果を報告する。

1　はじめに

英語教育において、文章の一部を空所化して適切な語を補わせる穴埋め問題は、語彙知識、文法知識、文脈情報を統合して正解を導く能力を測る統合的テストとして重要視されている [1]。この問題で学習者の読解力を適切に評価するには、空所周辺の局所的な情報だけではなく、パッセージ全体の内容的つながりに基づいた推論を経て正解に辿り着ける問題（Long Inference 型、以後 Long 型と呼ぶ）が不可欠である [2, 3]。そのような問題を作成するには、文脈依存度に基づいて削除する語を合理的に選定する必要がある [4]。

こうした合理的削除を大規模かつ効率的に実現する手段として、近年の自然言語処理分野では、言語モデルを用いた英語学習向けの問題自動生成の研究が進展している [5, 6]。特に、Ondov ら [7] は、長文穴埋め問題の自動生成の手法として、マスク言語モデル（MLM）の予測確率を利用した手法（nCloze）を提案し、文脈推論を必要とする問題の生成に取り組んだ。しかし、生成された問題は、教師作成テストとの受験者得点の相関を見る方法や、生成された問題の間で同じ測定結果が得られるかの観点で評価された一方、本質的に重要な Long 型の問題が作成されたかの直接的な評価は行われていない。

そこで、本研究ではまず、英語長文穴埋め問題において本質的に必要な Long 型の問題の生成量を直接的に評価方法を確立する。このために、LLM-as-a-Judge [8] の手法を用い、作成された穴埋め問題のカテゴリを分類する手法を考案する（§3）。提案する分類手法で既存手法（nCloze） を分析した結果（図 1）、nCloze では比較的高い割合で Long 型（図の Inference-Probable と Inference-Ambiguous 相当）の作成に成功している一方で、パッセージ中の単語照合で単純に解ける Matching 型の問題も高い割合で生成してしまうことが明らかになった。また、空欄を含む 1 文のみで解ける問題 (Short) や、本質的に選択肢間に曖昧性が残る問題 (Unsolvable) も一定量生成されることが分かるなど、より細かい粒度での性質分析が可能となった。

この結果を受けて、本研究ではさらに nCloze の空欄箇所選定を改善するために、2 つの観点から空欄箇所選定に用いる数式の改良を試みた（§4）。

実験により、提案する改善手法が Matching 型の問題の大部分を削減し、Long 型の問題を増加させることが確認された。

2　実験データ

本研究では、評価および実験用データとして CLOTH データセット [9] を用いる。これは、中学校および高校の英語入学試験から収集された、英語教師によって作成された穴埋め問題データセットである。データセット全体は 7,131 件のパッセージと 99,433 問の問題から構成され、各パッセージには平均して約 14 箇所の空欄が含まれており、各空欄は 4 肢選択式問題（正解となる元の単語 1 つと、誤答選択肢 3 つ）として構成されている。実験では、このデータのうち、Ondov ら [7] が実験で利用していたのと同一の 18 パッセージを利用する。

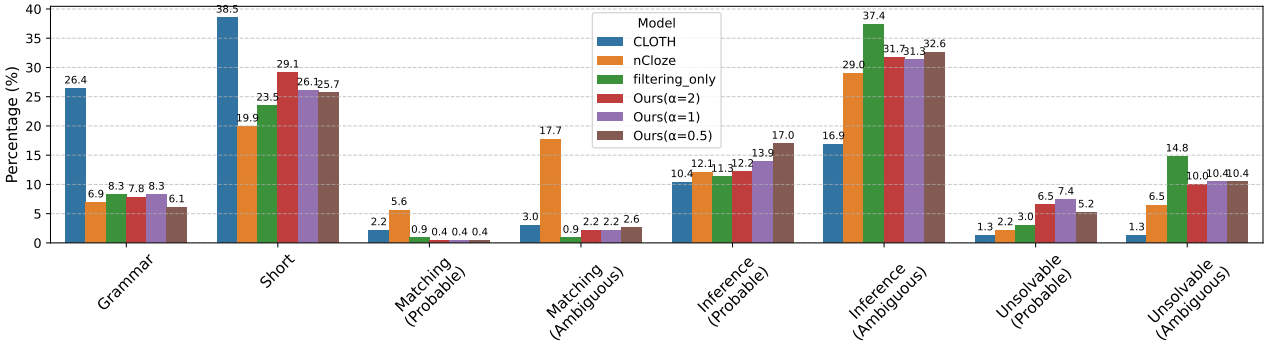

図 1　LLM-as-a-Judge を用いた長文穴埋め問題の自動評価結果

3　LLM-as-a-Judge による分類

正解の導出に必要な情報範囲や推論の種類に基づいて作成された問題をカテゴリ分類するため、本研究では大規模言語モデル（gpt-5-2025-08-07）を用いた自動分類手法 (LLM-as-a-Judge) を構築する。近年の研究では、高度な推論能力を持つ LLM が人間の専門家と極めて高い相関を持つ評価を行えることが示されており [8]、教育ドメインにおいても活用が見られる [10]。

3.1　方法（分類手順）

各空欄に対し、以下の 3 つの分類タスクを順に行い、その判定結果の組み合わせで最終的なカテゴリを決定する（プロンプトの詳細は付録 A を参照）。

まず、空欄が文法規則やイディオムなどの定型表現のみで正解可能かを判定し、これに該当する場合はカテゴリを **Grammar** とする（Step 1）。

次に、空欄を含む 1 文のみを入力とし、その情報だけで正解を 1 つに絞り込めるかを判定する（Step 2）。この判断には、判断を行う者の読解力や主観による揺れが生じうることを考慮し、次の 3 段階で判定を行った。1 文内の手がかりだけで、**Short**: 高い確信度で正解を 1 つに絞り込める、**Probable**: 中程度の確信度で正解を 1 つに絞り込める、**Ambiguous**: 手がかり不足で正解を 1 つに絞り込めない。このうち、Probable と Ambiguous については、文脈情報により確信度が向上する可能性のある事例として、次の判定を続けて行う。

最後のステップ (Step 3) では、パッセージ全文を入力とし、文脈情報を用いた解決可能性に基づき次のいずれかに分類する。**Matching**: 文脈上に正解と同一（または同義）の単語が存在し、その照合のみで解ける[1)]。**Inference**: 文脈上の推論によって解答の確信度が高まる。**Unsolvable**: 文脈が解答に寄与する情報を与えない。

以上の手順により、各空欄は図 1 に示す 8 つのカテゴリに分類される。本研究では、Step 3 で Inference と判定されたもの (Inference-Probable および Inference-Ambiguous) を Long 型に相当する良質な文脈推論型問題と位置づける。一方で、Unsolvable-Ambiguous と判定されたものについては、空欄を含む 1 文からも文脈からも解答を絞り込むための十分な情報を得られない不良問題とみなす。

なお、評価の安定性を確保するため、同一の問題に対して 5 回試行を行い、多数決によって決定した（同数の場合は同率 1 位からランダム選択）。

3.2　人手分類との整合性検証

提案手法の信頼性を検証するため、CLOTH データセットからランダムに抽出した、実験用データとは異なる 7 つのパッセージ（計 100 問の空欄）に対し、著者による人手分類と LLM による自動分類の結果を比較した。図 2 にその混同行列を示す。結果として、Grammar と Short のカテゴリでは高い整合性が確認された。人間と LLM の間の揺れは、事前の観察から予測される通り、Short（高い確信度で 1 文で解ける）と Probable （中程度の確信度で 1 文で解ける）の間に集中しており、主観的なゆらぎの範囲であることが観察できた。また、LLM は人間と比べて Probable よりも Short の判定に偏る傾向が見られた。この結果は、LLM が人間よりも正確な生起確率の知識を持ち、より保守的に文脈推論が不要であると判定した結果だと推測される。以上の結果

1)　ただし、誤答選択肢のいずれかにおいて、同じく文脈上に同一/同義の単語が見られるものについては、解答の曖昧性が生じるため Matching とは判定しない。

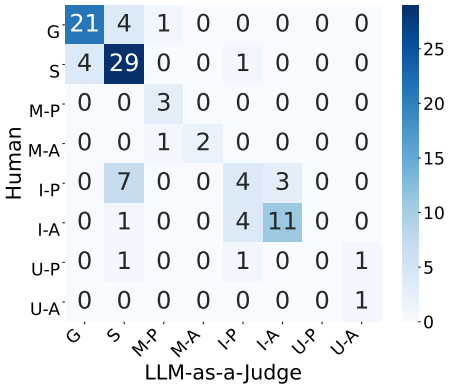

図 2　LLM の問題分類結果と著者らの結果との混同行列。G は Grammar、S は Short、M は Matching、I は Inference、U は Unsolvable、P は Probable、A は Ambiguous を表す。

から、本分類方法により LLM が Inference と判定した事例については「文脈依存性が高い問題」の信頼性の高い下限値を示すものと考えることができる。

4　空欄箇所選定の評価式の改良

本研究では、Ondov ら [7] の既存手法（以下、nCloze）を用いて問題生成タスクを再実験し[2)]、前節で述べた LLM を用いた自動分類フレームワークによる詳細な分析を行った。実験にあたっては、各パッセージに対して、CLOTH データセットにおいて人間が作成した空欄箇所と同数の問題を生成させた。分析の結果（図 1）、nCloze は文脈推論を必要とする Inference 型の問題を比較的多く生成できる一方で、本文中の単語を単純に照合すれば解ける Matching 型の問題を過剰に生成する傾向にあることが示された。また、正解を 1 つに絞り込めない不良設定問題（Unsolvable 型）も一定数観察された。そこで我々は、より高品質な文脈推論型問題を生成するために、nCloze に対して 2 点の改良を実施する。本節では、まずベースラインとなる nCloze の手法について概説し、その後、我々が提案する 2 つの改善手法について詳述する。

4.1　既存手法: nCloze

nCloze は、MLM の予測確率を利用して、英文 1 文としての自然さと文脈的な不整合性を両立させる誤答が含まれる問題を生成する手法である。この手法は、入力として問題を作成する対象となるパッセージを受け取り、(1) パッセージ内の適切な単語を正解として空所化したうえで、(2) それらに対して、それぞれ 3 つの誤答選択肢を生成することで、4 択形式の穴埋め問題を出力する。

空欄箇所の選択　与えられたパッセージ中のすべての単語位置を空欄候補とし、それぞれの位置で誤答として適した単語が存在するかを評価する。まず、ある単語位置をマスクし、1 文またはパッセージ全体を文脈として MLM により単語予測確率を計算する。ここで、その位置で、ある単語が誤答候補として適格かを評価するために、以下のスコア関数 $s(d)$ を用いる。なお、原著論文中の表記と公開されている実装には差異が見られるため、本稿では公開されている実装コードに基づいた数式で説明する。

$$s(d) = \ln P_{\mathrm{s}}(d) - \alpha \cdot \ln P_{\mathrm{p}}(d) \tag{1}$$

ここで、d は誤答候補単語である。$P_{\mathrm{s}}(d)$ は空欄を含む「1 文のみ」をモデルに入力した際の単語 d の予測確率で、Plausibility と名付けられている。これは、その誤答選択肢が埋められたときの 1 文としての自然さを表しており、確率が高いほど局所的にはもっともらしく見える良い誤答選択肢であることを示す。$P_{\mathrm{p}}(d)$ は「パッセージ全体」を入力した際の予測確率で、Incorrectness （文脈全体での不適切さ）と名付けられている。実際には、この確率が「低いほど」文脈上不自然となる「適切な誤答選択肢」を表す指標となる（低いほど良い）。α は文脈確率に対する重みのハイパーパラメータであり、探索の結果、$\alpha = 0.3$ が採用されている。この式に基づき、nCloze では「1 文としては自然（P_{s} が高い）だが、文脈全体で見ると不自然（P_{p} が低い）な誤答候補」ほど、高いスコアが与えられる。

最終的に、このスコアが高い上位 k(=32) 個の誤答候補 D を用いてスコアの総和を取り、これをその位置の空欄候補としての評価値 $S(D)$ とする。

$$S(D) = \sum_{d \in D} s(d) \tag{2}$$

つまり、誤答候補として的確な単語が多く見つかる位置ほど、適切な空欄箇所とみなされる。

誤答選択肢の生成　本研究では、誤答選択肢の探索アルゴリズムについては nCloze に変更を加えず、同一の方法を利用するため、ここでは概要のみを説明する。nCloze では、式 2 の評価軸に加え、式の第 2 項として「選択肢間の単語ベクトルの非類似度（Distinctiveness）」を考慮している。これは、正解の選択肢および誤答選択肢の間で意味が似通った選択肢が選ばれるのを防ぐためである。具体的な生成プ

2）　https://github.com/bionlp-nlm/ncloze に公開の実装を用いた。

ロセスとしては、まず式 1 の上位スコアを持つ単語を候補プールとして保持しておき、そこから焼きなまし法を用いて、スコアの合計が最大となる 3 つの誤答の組み合わせを探索する。

4.2 単語類似度による Matching の抑制

nCloze による Matching 型問題の過剰な生成を抑制するため、新たに空欄箇所の決定プロセスにフィルタリング処理を導入する。空欄候補位置の単語 w を中心として、前後 W トークン（本実験では $W = 500$）に含まれるすべての単語セット C_w に対して次の 2 つの処理を行う。(1) 完全一致: w と同一の文字列が C_w 内に存在する場合、その箇所を除外する。(2) 意味的類似（Semantic Similarity）: 単語埋め込みベクトルを用いて、文脈内に同義語がある場合を除外する。(2) には、spaCy の en_core_web_lg モデル[3)]により得られる単語ベクトルを用い、候補単語 w と文脈語 $c \in C_w$ とのコサイン類似度 $\cos(\vec{w}, \vec{c})$ を計算する。この最大値が閾値 θ（本実験では $\theta = 0.8$）を超えた場合、同義語がパッセージ内に存在するとみなし、その箇所を除外する。

4.3 正解選択肢との確率比による評価

正解の選択肢と一部の誤答選択肢の間に、本質的な曖昧性の残る問題が生成されてしまう課題に対処するために、誤答候補単語 d の質を「正解単語 a との相対的な比較」によって評価する新たなスコア関数を導入する。具体的には、空欄箇所における誤答選択肢の単純な生起確率を用いる代わりに、正解選択肢の生起確率との比を考える。

$$\text{Plausibility}(d, a) = f\left(\frac{P_s(d)}{P_s(a)}\right) \tag{3}$$

$$\text{Incorrectness}(d, a) = 1 - f\left(\frac{P_p(d)}{P_p(a)}\right) \tag{4}$$

$$f(x) = \sigma\left(10 \cdot (x - 0.5)\right) = \frac{1}{1 + \exp\{-10 \cdot (x - 0.5)\}} \tag{5}$$

ここで、$f(x)$ は S 字カーブの中心が $x = 0.5$ の地点にあり、$x = 0$, $x = 1$ で値がそれぞれ 0 と 1 に飽和するシグモイド曲線である。つまり、新たな Plausibility は、空欄を含む 1 文を文脈として与えたときの d の生起確率が a の生起確率と同等かそれ以上の場合に良いとみなし、誤答単語の生起確率が相対的に低いほど悪いとみなす。新たな Incorrectness は、パッセージ全体を文脈として与えたときの d の生起確率が a の生起確率と同等かそれ以上の場合にスコアが 0 に近くなり、誤答選択肢の確率が相対的に低いほど良い選択肢とみなす。

最終的にこれらのスコアを用いて (1) 式が次のように置き換えられる。

$$s(d, a) = \ln f\left(\frac{P_s(d)}{P_s(a)}\right) + \alpha \cdot \ln\left(1 - f\left(\frac{P_p(d)}{P_p(a)}\right)\right) \tag{6}$$

5 実験結果

単語類似度による抑制の効果 図 1 の filtering_only は、提案する Matching 型抑制の手法のみを nCloze に適用した場合の結果を示している。結果として、我々の期待通り Matching 型の問題は 23.3% から 1.8% に 21.5% 抑制され、それに伴って Long 型の問題が 7.6% 増加した。一方で、本質的に正解の曖昧性が残る Unsolvable-Ambiguous の問題が顕著に増加したほか、Grammar と Short の問題も微増する結果となった。

確率比による評価導入の効果 図 1 の Ours ($\alpha \in \{0.5, 1, 2\}$) は、上記フィルタリングに加えて、我々が提案する式 6 を採用した場合の結果である。結果として、いずれの α の場合も filtering_only に比べて、Ambiguous 型（1 文のみでは解けない）に対して Probable 型（1 文のみで中程度の確信度で解ける）を増加させる傾向にあったが、Unsolved 型の大幅な抑制には至らなかった。

全ての結果の中で Long 型の問題が最も多く作成されたのは提案手法のうち $\alpha = 0.5$ を用いた場合であった (49.6%) が、長文読解能力を測定する上で最も適切なカテゴリと考えられる Inference-Ambiguous が最も多いのは filtering_only であった。

6 おわりに

本稿では、英語長文穴埋め問題において、生成された問題の自動分類に基づく評価手法を提案し、この分析結果を用いて既存手法の改善を試みた。提案した分析手法は、生成された問題の選別にも利用可能なため、Long 型の問題をできる限り生成した後に、不適切な問題を除去するといった応用にも利用可能である。今後の方向性として、不良設定問題のさらなる抑制方法の検討や、特定のパッセージ内で本質的に作成できる Long 型問題の上限値を推定することで、長文読解問題により適切なパッセージ自体を選択することなどを検討している。

3) https://spacy.io/models/en

謝辞

本研究は JSPS 科研費 JP25K00470 の助成を受けたものです.

参考文献

[1] John W Oller, Jr. CLOZE TESTS OF SECOND LANGUAGE PROFICIENCY AND WHAT THEY MEASURE[1]. **Lang. Learn.**, Vol. 23, No. 1, pp. 105–118, June 1973.

[2] J Charles Alderson. The cloze procedure and proficiency in english as a foreign language. **TESOL Q.**, Vol. 13, No. 2, p. 219, June 1979.

[3] John Jonz. Another turn in the conversation: What does cloze measure? **TESOL Q.**, Vol. 24, No. 1, p. 61, March 1990.

[4] Lyle F Bachman. Performance on cloze tests with fixed-ratio and rational deletions. **TESOL Q.**, Vol. 19, No. 3, p. 535, September 1985.

[5] Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. CDGP: Automatic cloze distractor generation based on pre-trained language model. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 5835–5840, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[6] Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. Distractor generation based on Text2Text language models with pseudo Kullback-Leibler divergence regulation. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 12477–12491, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Brian Ondov, Kush Attal, and Dina Demner-Fushman. Pedagogically aligned objectives create reliable automatic cloze tests. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 3961–3972, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[9] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale cloze test dataset created by teachers. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2344–2356, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Nisarg Parikh, Nigel Fernandez, Alexander Scarlatos, Simon Woodhead, and Andrew Lan. Lookalike: Consistent distractor generation in math mcqs, 2025.

A　LLM 評価で用いたプロンプト

```
% \begin{Verbatim}[breaklines]
SYSTEM_PROMPT = """
You are an expert linguistic analyst. Your task is to
     evaluate the "Solvability" of a cloze question
     through a strictly defined 3-step process.
You must respond **only** with a single JSON object.
### Output Format (Strictly follow this order)
{
    "passage_id": "...",
    "question_number": "...",
    "reasoning": "Step 1: [Grammar Analysis]. Step 2: [
     Local Sentence Analysis]. Step 3: [Long Context &
     Lexical Analysis]. -> [Conclusion]. (IMPORTANT:
     Perform the analysis logically BEFORE deciding the
     category.)",
    "step1_grammar": boolean,
    "step2_local_status": "Solved" | "Probable" | "
     Ambiguous",
    "step3_resolution": "Matching" | "Inference" | "None"
      | "N/A",
    "final_category": "CATEGORY_NAME"
}
(CATEGORY_NAME must be one of: Grammar, Single-Sentence-
     Solved, Long-Context-Matching, Long-Context-
     Inference, Probable, Ambiguous)
---
### Analysis Process (Follow these steps)
**Step 1: Grammar Check**
* **Definition:** Determine if the question is about **
     grammar usage**, involving tense, preposition usage,
      active/passive voices, subjunctive mood and so on.
* If the question fits this definition -> Final Category
     is **Grammar**.
**Step 2: Single Sentence Analysis (Local Context)**
* Read **ONLY** the single sentence containing the blank.
* **Status: Solved (High Confidence):** Clues within the
     sentence are sufficient to pinpoint the answer. (
     Final Category: **Single-Sentence-Solved**).
* **Status: Probable (Medium Confidence):** One option is
      likely, but context *could* change it.
* **Status: Ambiguous (Low Confidence):** Impossible to
     determine without context.
**Step 3: Long Context Analysis (Full Passage)**
* *Perform only if Step 2 was Probable or Ambiguous.*
* **CRITICAL CHECK: Strict Lexical Analysis**
    * Check if the **Correct Answer** appears in the text
     .
    * **Definition of "Appears" (Strict Matching):**
        1. **Exact match**
        2. **Lemma match:** (e.g., "go" matches "went").
        3. **Singular/Plural match:** (e.g., "apple"
     matches "apples").
        4. **Compound/Morphological inclusion (Headword
     Match):**
            The correct word is contained within a
     compound word in the text AND retains its core
     meaning.
            * **YES (Matching):** "light" matches "
     sunlight" (sunlight IS light). "ball" matches "
     football" (football IS a ball).
            * **NO (Not Matching):** "rain" in "rainbow" (
     Different concept). "book" in "notebook" (Related
     but distinct objects).
    * **Rule A (Matching):** Correct answer appears (
     valid match), and Distractors do NOT appear. -> **
     Long-Context-Matching**.
    * **Rule B (Inference):** Correct answer appears, BUT
      Distractors ALSO appear (requiring context to
     distinguish). -> **Long-Context-Inference**.
    * **Rule C (Inference):** Correct answer does NOT
     appear. -> **Long-Context-Inference**.
---
### Few-Shot Examples (Reference)
**Example 1: Grammar**
*Context:* "... She happened to see on the desk a half-
     opened notebook, which __ : " In order to keep the
     secretaries in high spirits, the company has decided
      that every Monday morning a bunch of fresh flowers
     should be put on each secretarys desk." ..."
*Choices:* (A) said, (B) written, (C) printed, (D) signed
*Analysis:*
{
     (スペースの都合上省略)
}
**Example 2: Short (Single-Sentence-Solved)**
*Context:* "... Monday was the first day she went to work
     , so she was very __ and arrived early. ..."
*Choices:* (A) depressed, (B) encouraged, (C) excited, (D
     ) surprised
*Analysis:*
{
     (スペースの都合上省略)
}
**Example 3: Matching (Ambiguous in Local -> Matching in
     Global)**
*Context:* "... She was surprised to find a bunch of __
     on it. They were fresh. She smelled them and they
     were sweet. ... " Somebody has sent me flowers the
     very first day!" she thought happily." ..."
*Choices:* (A) keys, (B) grapes, (C) flowers, (D) bananas
*Analysis:*
{
     (スペースの都合上省略)
}
**Example 4: Inference (Ambiguous in Local -> Inference
     in Global)**
*Context:* "... She pushed the door open and found nobody
      there. " I am the __ to arrive." She thought and
     came to her desk. ..."
*Choices:* (A) last, (B) second, (C) third, (D) first
*Analysis:*
{
    "reasoning": "Step 1: Not grammar. Step 2: 'I am the
     ___ to arrive' is Ambiguous; any ordinal number or '
     last' could fit grammatically. Step 3: The clue is
     in the preceding sentence: 'found nobody there'.
     This logically implies she is the 'first' person.
     Even if the word 'first' appears elsewhere in the
     text (e.g., 'first day'), solving this blank
     requires the logical deduction from 'nobody there',
     not just finding the word. Thus, it is Inference.",
    "step1_grammar": false,
    "step2_local_status": "Ambiguous",
    "step3_resolution": "Inference",
    "final_category": "Long-Context-Inference"
}
"""
...
```