PBIG-DATA: 専門家による評価スコア付き製品アイデアデータセット

広田航 [1] 谷口友紀 [2] 大熊智子 [2] Chung-Chi Chen[3] 高橋洸丞 [1] 進藤尚希 [1,4]
有馬幸介 [1] 朝倉卓人 [1,5,6,7] 近江崇宏 [1] 石垣達也 [3]
[1] ストックマーク株式会社 [2] 旭化成株式会社 [3] 産業技術総合研究所
[4] 電気通信大学 [5] 国立情報学研究所 [6] 東京大学 [7] 理化学研究所
wataru.hirota@stockmark.co.jp ishigaki.tatsuya@aist.go.jp

概要

本論文では，専門家による評価スコアを付与した新規製品アイデアからなる大規模データセット「PBIG-DATA」を提案する. PBIG-DATA はLLMが生成したアイデアに対して技術・ビジネス分野の専門家が付与した約 3,000 件の多角的な評価スコアを含む．評価スコアの分析および LLM-as-a-Judge の性能評価を通して製品アイデアの自動生成における評価者間のスコアの揺れや個人の評価基準の一貫性といったタスクの性質を明らかにする.

1 はじめに

特許文献などの既存技術に関する説明文書を入力とした製品アイデア生成は, 技術活用を目指す学術界および産業界の双方において注目されている [1, 2]. 従来，アイデアの良し悪しを判断するための評価スコアが付与された大規模なデータセットが存在せず, 他の創造的なアイデア生成タスクで報告されているような, 評価スコアの主観性などのタスクの特徴分析が行われてこなかった [3, 4].

そこで我々は, 専門家による評価スコアが付与された製品アイデアデータセット PBIG-DATA を提案し，本タスクの性質を分析する．PBIG-DATA には LLM が生成した製品アイデア約 300 件と, それらに対して技術評価者およびビジネス評価者によって人手付与された評価スコア約 3,000 件が含まれる．アイデアの評価は具体性, 技術的妥当性, 革新性, 競争優位性, ニーズの妥当性, 市場規模という複数の評価軸によって行われた.

本論文ではこのデータセットの分析から分かった 2 つの重要な性質を報告する．第一に, 評価者間でスコアの分布が多様で主観性が高い評価タスクであるという点である．これは製品アイデア生成という創造的なテキスト生成タスクの評価が，単一の正解が定まらない本質的に多様な視点を要するものであることを示唆する. 第二に, 同一評価者の評価の一貫性は高いという点である. 本研究では, パーソナライズを行った LLM-as-a-Judge モデルがパーソナライズを行っていないモデルを比較して, その評価者とのスコアの一致率が高まることを示す. これは, このような創造的なアイデア生成タスクであるとしても各評価者はそれぞれ一貫した評価方針を持つことを示唆する．以上の結果より，本稿では製品アイデア生成の評価タスクにおいて「誰か評価するか」のモデリングが重要であると主張する.

2 関連研究

創造的なテキスト生成タスクにおいて，評価データセットの構築は活発に行われている．例えばストーリー生成やブレインストーミングタスクなどにおいては，人間による評価付きデータセットが提案され，モデルの創造性評価に活用されている [3, 4]. しかし，既存技術を起点とした製品アイデア生成（製品アイデア生成）は，単なる自由発想とは異なり, 入力となる技術の仕様を満たしつつ（技術的実現性），市場における価値（ビジネス性）を両立させる高度な発想が求められる．既存のデータセットはこのような「技術シーズに基づく制約付き創造性」を扱っておらず，また技術・ビジネスの専門家による多角的な評価スコアが含まれていないため，製品アイデア生成の評価基盤としては不十分であった.

生成されたアイデアの評価コストを削減するため，LLM-as-a-Judge による自動評価が注目されている．実際に，LiveIdeaBench [5] や IDEA-Bench [6] といったベンチマークでは，LLM を用いた自動評価

表 1 PBIG-DATA の統計.

カテゴリ	アノテーター数	特許数	アイデア数	評価数
NLP	12	46	100	1,055
CS	11	48	97	984
MC	4	22	110	1,070

指標が提案されている．LLM-as-a-Judge は多様なタスクで有力な選択肢であるが，一方で寛大化バイアスやプロンプトへの感度といったバイアスも指摘されている [4]. したがって，製品アイデア生成のような主観性が高く専門知識を要するテキスト生成タスクにおいて LLM による自動評価を採用するには，まずその評価能力自体を検証（メタ評価）する必要がある．これまで，このメタ評価を行うための信頼できる「専門家による正解評価データ」が存在しなかったことが，自動評価技術の導入を阻む障壁となっていた.

3 PBIG-DATA データセット

提案データセットには製品アイデアと専門家が付与した評価スコアが含まれる．本節でそれぞれの収集手法を説明する.

3.1 製品アイデアの収集

PBIG-DATA には Product Business Idea Generation (共通タスク PBIG) [1] で提出された製品アイデアが含まれる. これらの製品アイデアは, 英語の特許文献の全文を入力とし, その特許技術を活用した製品アイデアを LLM に生成させたものである．入力特許は, 1) 自然言語処理 (NLP), 2) コンピュータサイエンス (CS), 3) 素材化学 (MatChem) の 3 分野から選ばれたものを使用する．各入力特許には, 構造化されたメタデータ (タイトル, 特許公開番号, 要約, 請求項, 説明) が含まれる．アイデアの出力はタイトル (最大 100 文字), 製品説明, 実装方法, 差別化要素 (それぞれ最大 300 文字) から構成される. 各項目は英語で記載される. 付録 A にアイデアの例を示す.

3.2 評価スコアのアノテーション

本データセットの評価基準は, 共通タスク PBIG で用いられたルーブリック [1] に従う. 実際のルーブリックは付録 B を参照されたい．本研究では評価軸を技術的な評価軸 (技術的妥当性, 革新性, 競争優位性) とビジネス的な評価軸 (ニーズの妥当性, 市場規模) に分類する. 同様にアノテーターも専門性に応じて技術評価者とビジネス評価者の 2 チームに分け, 技術的な評価軸からは技術評価者が, ビジネス的な評価軸からはビジネス評価者がそれぞれ評価を行った. 表 1 に各カテゴリにおけるアノテーター数, 特許数, アイデア数, 評価数を示す．同一のアイデア・評価軸に対し複数人のビジネス評価者と複数人の技術評価者がそれぞれ評価を行った.

表 2 スコア評価における評価者間一致度 (Krippendorff の α 係数). サンプルサイズが 10 未満の評価軸 (*) は計算対象外とした.

評価軸	NLP	CS	MatChem
具体性	0.06	-0.11	0.04
技術的妥当性	-0.03	-0.40	-0.28
革新性	0.33	0.47	0.46
競争優位性	-0.08	0.24	-0.02
ニーズの妥当性 (B2B)	-0.23	0.02	0.05
ニーズの妥当性 (B2C)	*	-0.20	-0.22
市場規模 (B2B)	0.48	-0.31	0.08
市場規模 (B2C)	*	0.01	0.00

予備実験において, 曖昧なアイデアや実現不可能なアイデアはその他の評価軸から意味のある評価ができないことがわかった．そこで，本研究では次のような段階的な評価手順を導入した．技術評価者は, 具体性スコアが 2 より大きい場合のみ, 技術的妥当性を評価する．具体性と技術的妥当性のスコアがそれぞれ 2 と 1 より大きい場合のみ, さらに革新性と競争優位性を評価する. ビジネス評価者は, 具体性スコアが 2 より大きい場合のみ, ニーズの妥当性と市場規模の評価に進む.

4 分析

本節では製品アイデアの評価の多様性と主観性を分析する. 評価スコアに関して, 我々は以下の 2 つの仮説を立て, 検証した: (1) 生成されたアイデアの品質評価において評価者には暗黙のバイアスが存在し, **評価者間の一致度は低くなる**. (2) 一方, 各評価者は一貫した内部評価基準を持っており, **同じ評価者のスコアの一貫性は高くなる**.

4.1 評価者間の一致度

表 2 に Krippendorff の α 係数 [7] を示す．Krippendorff の α 係数は評価者の一致度を表す指標であり, –1 から 1 の範囲の値を取る. α が 1 に近いと一致度が高く, 0 に近いと一致度が低く, –1 に近いと不一致が生じていると解釈できる．表 2 において多くの α がゼロに近いか, あるいは負の値となっており, 評価

図 1 LLM-as-a-Judge のスコアと評価者の評価スコアとの一致度．各プロットの X 軸は評価履歴に含まれる事例数 (n-shot) を，Y 軸は Krippendorff の α 係数を表す．緑色，青色，赤色の線はそれぞれ `zero shot`, `few shot (diff)`, `few shot (same)` の結果を表す．

者間の一致度は低いことがわかる．評価軸別にみると，革新性の評価は全体的に高い一致度を示す一方，技術的妥当性の評価は低い一致度が得られた．この理由の 1 つとして，技術的妥当性のように特許と製品アイデアの各所を詳細に評価するような評価軸は，革新性 のような総合的な評価軸よりも判断の合意が難しいという点が考えられる．

製品アイデア生成のような創造的なタスクの場合，評価者の意見の相違はタスク固有の主観性に由来することが多く [4]，本質的に評価者間一致度が低くなるケースが報告されている．本タスクにおいても低い Krippendorff の α 係数は評価タスクの特徴を表すものと捉えられ，必ずしもアノテーションスコアの品質が低いことを意味するものではない [8]．

4.2 評価者内の一貫性

次に「各評価者が内部に一貫した評価基準を持つ」という仮説を検証するために，評価者ごとにパーソナライズされた LLM-as-a-Judge モデルとパーソナライズを行わない LLM-as-a-Judge モデルの自動評価性能を比較する．もしその評価者の評価事例によってパーソナライズされたモデルがパーソナライズされていないモデルと比べて高い自動評価性能を示す場合，それは評価者の評価事例には一貫した内部基準が反映されていることを意味する．

4.2.1 実験設定

モデルと推論方法 本実験では指示チューニングがされた 4 つの Qwen3 [9] モデルを使用する．1 つは推論モデル (`30B-A3B-Thinking-2507`) であり，残りの 3 つは非推論モデル (`4B-Instruct-2507`, `30B-A3B-Instruct-2507`, `235B-A22B-Instruct-2507`) である．本実験では異なるランダムシードで 3 回の推論を実行し，予測されたスコアの多数決により最終的なスコアを決定する．Dong らの手法 [10] に従い，LLM には予測と共に 0 から 100 の間で信頼度を出力させ，信頼度が 80 未満の推論は棄却した．

プロンプトとパーソナライズ LLM-as-a-Judge のプロンプト手法として `zero shot`, `few shot (diff)`, `few shot (same)` の 3 つを用いる．(図 2). `zero shot` は，タスクの説明のみをプロンプトに含める．`few shot (diff)` はタスクの説明に加え他の評価者からランダムにサンプルした評価事例の履歴をプロンプトに加える．`few shot (same)` はパーソナライズを行う設定であり，`few shot (diff)` と同様に評価事例の履歴をプロンプトに加えるが，その評価事例は評価対象の評価者のものを用いる．

4.2.2 結果

図 1 に LLM-as-a-Judge のスコアと評価者のスコアの一致度 (Krippendorff の α 係数) を示す．多くの評

LLM-as-a-Judge Prompt Template

```
You are given a pair consisting of a patent
and a product idea based on that patent. Your
 task is to evaluate the idea following the
given instruction. First, you will receive a
detailed instruction. If the setting is few-
shot, several examples of patents, ideas, and
 scores are also provided.Finally, you will
be given a new patent and idea to evaluate.

## Instruction
<Instruction text here>

## Examples (only in few-shot setting)
<Example 1: patent, idea, and score>
...
<Example N: patent, idea, and score>

## Input
<Patent and idea to be scored>

## Output format
Return a single line of valid JSON in this
format: {"score": <number>, "reason": "<brief
 reason>", "confidence": <integer between 0
and 100>}
```

図 2 LLM-as-a-Judge モデルのプロンプトテンプレート. (<...>) はプレースホルダーを示す.

価軸・モデルにおいて, パーソナライズされた LLM-as-a-Judge (`few shot (same)`) が, `zero shot` や同じ事例数の `few shot (diff)` よりも高い一致度を示す. これは, 各評価者は一貫した評価基準がある一方で, それが評価者間で異なることを示唆する. 一方, 興味深いことに, 30B 以下の小型モデルでは `few shot (same)` が `few shot (diff)` よりも低い α を示す場合があった. これは, 事例から個々の評価者の傾向を理解する能力が LLM の能力に依存していることを示唆する.

4.3 自動評価の理由からみる評価者の内部一貫性

最後に, 評価者間の類似性と LLM-as-a-Judge が生成する評価の理由の類似性を比較することで, 評価者の内部評価ロジックの一貫性を間接的に検証する. もしパーソナライズされた LLM-as-a-Judge が出力した理由の類似度が評価間一致度が高い評価者ペアほど高い場合, それは単なる偶然の一致ではなく, 根底にある評価ロジックを共有していることが示唆される.

図 3 に共通したアイデアを評価した評価者の全ての組み合わせにおける評価者間一致度と LLM-as-a-Judge モデルが生成した評価の理由の類似度の散布図を示す. 評価の理由は図 1 で最も高い自動評価性能を示した `Qwen3-235B-A22B-Instruct-2507` の出力を用いる. 評価の理由のテキスト埋め込みには `Qwen3-Embedding-8B`[1] を用いる.

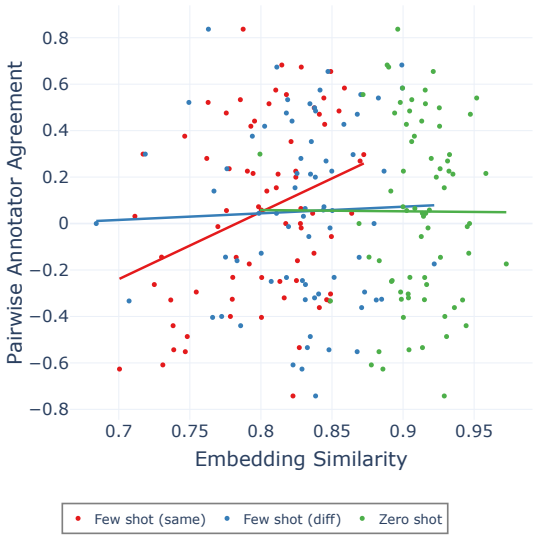

図 3 評価者間一致度 (Y 軸, Krippendorff の α 係数の平均値) と, パーソナライズ版 LLM-as-a-Judge が生成した根拠テキストの類似度 (X 軸, コサイン類似度の平均値) の関係. 各点は評価者のペアを表す. 赤, 青, 緑はそれぞれ `few shot (same)`, `few shot (diff)`, `zero shot` を示す. 色付きの線は各プロンプト手法における線形回帰を示す.

結果, `few shot (same)` でのみ明確な正の相関 ($r = 0.311$) が見られ, `few shot (diff)` ($r = 0.036$) および `zero shot` ($r = -0.003$) では明確な相関は確認されなかった. この結果は, より高い一致度を持つ評価者ペアの内部的な評価基準が部分的に収束した推論パターンを共有していることを示唆しており, 個々の評価者内における一貫した内部評価ロジックの間接的な証拠となる.

5 まとめ

本研究では, 専門家が評価した製品アイデアを含む公開データセット PBIG-DATA を構築し, その分析を通して, アイデア生成タスクにおける 1) 評価の主観性 2) 同一評価者での評価の一貫性という 2 つの重要な性質を明らかにした. 本研究が製品アイデア生成タスクのような創造的な生成タスクにおいて公平性と再現性を担保した評価システムを設計する研究の発展の一助となることを期待する.

1) `https://huggingface.co/Qwen/Qwen3-Embedding-8B`

参考文献

[1] Wataru Hirota, Chung-Chi Chen, Tomoko Ohkuma, Tomoki Taniguchi, and Tatsuya Ishigaki. Overview of PBIG shared task at AgentScen 2025: Product business idea generation from patents. In Chung-Chi Chen, Tatsuya Ishigaki, Sophia Ananiadou, and Hiroya Takamura, editors, *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning (AgentScen2025)*, pp. 35–42, Montreal, Canada, 16 August 2025.

[2] Chung-Chi Chen, Tatsuya Ishigaki, Sophia Ananiadou, and Hiroya Takamura, editors. *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning (AgentScen2025)*, Montreal, Canada, 16 August 2025.

[3] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[4] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 404–430, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics.

[5] Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Liveideabench: Evaluating llms' divergent thinking for scientific idea generation with minimal context, 2025.

[6] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing?, 2024.

[7] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[8] Barbara Plank. The ‘‘problem’’ of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[9] Qwen Team. Qwen3 technical report, 2025.

[10] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge?, 2024.

表 3　各評価軸のルーブリック.

Criterion	Description	Rubric
Specificity	Clarity and concreteness of the product description.	1. Cannot be read as coherent language. 2. Can be read as language, but the idea’ s meaning is barely conveyed. 3. One or more concrete products can be imagined. 4. A single concrete product can be clearly imagined.
Technical Validity	Implementation feasibility of the idea using the given patent.	1. The patented technology does not seem suitable for the use. 2. Building a prototype using the technology is challenging but possible. 3. A prototype could be built using the technology. 4. The technology can be applied to a production-level product.
Innovativeness	Degree of novelty and originality of the proposed solution.	1. A well-known application; lacks novelty. 2. Known use case of similar technology, but not yet fully explored. 3. A use case I hadn’ t thought of, but not particularly exciting. 4. Surprising and novel; strong originality. 5. Clearly innovative and potentially groundbreaking.
Competitive Advantage	Distinct benefits and advantages over existing solutions.	Two sub-questions are considered: (A) Is it difficult to imitate the idea using the technology? (B) Is the technology essential to the core of the idea? 1. Neither A nor B.　2. Only B.　3. Only A.　4. Both.
Need Validity	Relevance of the product to genuine user needs.	**B2B (0–3):** 0. Not a B2B product; 1. Both qualitative and quantitative returns are low; 2. Either quantitative (monetary) or qualitative (for corporate growth) returns are large; 3. Both qualitative and quantitative returns are large. **B2C (0–3):** 0. Not a B2C product; 1. Need is felt, but costs are low so most people do not seek a solution; 2. Need is felt with a certain level of cost; 3. Need is felt with significant cost.
Market Size	Number of potential users.	**B2B (0–3):** 0. Not a B2B product; 1 Niche, appeals to some companies; 2. Many companies acknowledge the issue; adoption depends on budget/systems; 3. Necessary for almost all ($\approx$ 80% or more) companies. **B2C (0–3):** 0. Not a B2C product; 1. Desired by some; not a daily necessity; 2. About 40–60% of people/households would want it; 3. 70–80% or more would want it; a daily necessity.

表 4　製品アイデアの例.

Patent No.	US1234567A
Title	Smart Sleep Assistant Pillow
Description	A pillow that monitors sleep posture and breathing using embedded sensors, providing real-time feedback and personalized adjustments to improve sleep quality.
Implementation	Applies the patented low-power sensor array technology to continuously track user movements and breathing, with data processed by an onboard microcontroller.
Differentiation	Unlike existing smart pillows, it combines posture correction with breathing monitoring, offering a holistic sleep improvement solution.

A　製品アイデアの例

表 4 に製品アイデアの例を示す.

B　ルーブリック

本研究で使用したルーブリックを表 3 に示す.