Overview of PBIG Shared Task at AgentScen 2025: Product Business Idea Generation from Patents

**Wataru Hirota,[1] Tomoko Ohkuma,[2] Tomoki Taniguchi,[2]
Chung-Chi Chen,[3] Tatsuya Ishigaki[3]**

[1]Stockmark, [2]Asahi Kasei Corporation, [3]AIST
wataru.hirota@stockmark.co.jp, ishigaki.tatsuya@aist.go.jp
okuma.td, taniguchi.tcr@om.asahi-kasei.co.jp

Abstract

This paper provides an overview of the shared task *Product Business Idea Generation from Patents (PBIG)*, held as part of the AgentScen2025 workshop at IJCAI2025. The task challenges participants to generate practical and innovative product business ideas based on real patent documents, under the constraint that the proposed product must be feasible to launch within three years. Participants were required to generate four textual components for each patent: product title, product description, implementation, and differentiation. The evaluation was conducted via pairwise comparisons using both large language models (LLMs) and human annotators across multiple criteria including technical validity, innovativeness, specificity, need validity, market size, and competitive advantage. This paper outlines the task setup, dataset structure, evaluation protocols, and discusses insights derived from participant submissions.

1 Introduction

Recent advances in large language models (LLMs) have enabled impressive capabilities in ideation tasks such as scientific discovery (Wang et al., 2024; Si et al., 2025; Lu et al., 2024; Keisuke et al., 2025) and future forecast (Ishigaki et al., 2022). However, the generation of viable business ideas grounded in real-world technologies remains a challenging and underexplored area. Unlike general language generation tasks, successful product ideation requires a combination of domain expertise, identification of unmet user needs, and the creative integration of novel technologies.

To address this challenge, we organized the shared task *Product Business Idea Generation from Patents (PBIG)* at the AgentScen-2025 workshop, co-located with IJCAI-2025. This task leverages patent documents as rich sources of technical knowledge and asks participants to propose product business ideas that could realistically be implemented within a short time frame (three years). The task aims to encourage natural language processing-based systems.[1] that are not only creative but also grounded in technical feasibility and market viability.

This overview paper presents the design, data, and evaluation protocols of the PBIG shared task. We also summarize the submitted systems, present results from automatic and human evaluations, and discuss open challenges and future directions.

2 Task Definition

This section describes the task, dataset, and evaluation protocols.

2.1 Task Overview

Participants were provided with real-world patent documents in text format, including both the abstract and the full description. Given this input, the task was to generate a product business idea that leverages the patented technology.

The proposed product must be something that could realistically be implemented and brought to market within three years. For each patent, systems were required to generate the following four textual components:

- **Product Title**: A concise name for the product (up to 100 characters).
- **Product Description**: A brief explanation covering the product's function, target users, user needs, and benefits (up to 300 characters).
- **Implementation**: A description of how the patented technology will be applied to realize the product (up to 300 characters).

[1] Idea generation models are not necessarily LLM-based but all submissions this time use LLM-based approaches.

- **Differentiation**: A description of what makes the product unique and how it stands out from existing solutions (up to 300 characters).

Participants were allowed to use any external knowledge sources in addition to the input patent, including other patents, web data, or APIs. Submissions were required to follow a structured JSON format specified by the organizers.

2.2 Dataset

The shared task dataset consisted of 150 patents sampled from the USPTO,[2] categorized into three technical domains: natural language processing (NLP), computer science, and material chemistry. Each patent was provided in a structured JSONL format with metadata (title, application/publication number and date), abstract, claims, and description fields. Additional materials, including patent PDFs and figure images, were also available in per-patent directories.

2.3 Evaluation

2.3.1 Overview

The submitted product business ideas were evaluated from six perspectives:

- **Technical Validity**: Is the idea technically feasible within three years?
- **Innovativeness**: Does the idea offer a novel solution to the demand?
- **Specificity**: Is the idea concrete and clearly articulated?
- **Need Validity**: Does the idea address an actual, well-defined user need?
- **Market Size**: Is the market large enough to justify the product?
- **Competitive Advantage**: What business advantage is gained by the idea?

Two types of annotators were involved in the evaluation process: domain **human experts** and **LLMs** (LLM-as-a-Judge).

2.3.2 Human Evaluation

Annotation Groups. Human experts were divided into two groups: the **technical group** and the **marketing group**. The technical group evaluated:

- Technical Validity
- Innovativeness
- Competitive Advantage

while the marketing group evaluated:

- Need Validity
- Market Size

For the NLP and Computer Science domains, manual annotation was conducted by NLP researchers from Stockmark and AIST. In the case of the **Material Chemistry** domain, experts from Asahi Kasei participated in both roles. All human evaluators are listed in the Acknowledgements section of this paper.

Sampling Ideas for Human Evaluation Due to the large number of submissions, we selected a subset of patents for human evaluation. For each selected patent, two annotators—one from the technical group and one from the marketing group—were assigned to evaluate each idea. In rare cases where assignment conflicts occurred, some ideas were evaluated by a single annotator.

Protocol Updates and Transition to Scoring Initial rounds of human evaluation were based on **pairwise comparisons**, in which annotators were shown two ideas and asked to judge which was better. However, agreement among annotators was low in this setting. To improve consistency, we transitioned to a **scoring-based protocol**, where each idea was assigned a numerical score for each criterion. Pairwise preferences could then be reconstructed by comparing scores from the same annotator.

We attach the full annotation guidelines in the appendix.

Pipeline-Based Annotation Protocol To handle low-quality or incomplete ideas and reduce annotation burden, we adopted a **pipeline-based evaluation strategy**. In this protocol, annotators sequentially evaluated each criterion and were allowed to skip subsequent criteria if earlier conditions were not met.

Technical Group Protocol:

1. **Specificity** is first scored on a 0–4 scale. If the score is 0–2 (i.e., the idea is too vague or unreadable), annotation stops.

[2] https://www.uspto.gov

2. If Specificity ≥ 3, the annotator proceeds to **Technical Validity** (0–4). If this score is ≤ 1, annotation also stops here.

3. If Technical Validity ≥ 2, **Innovativeness** is scored on a 0–5 scale.

4. **Competitive Advantage** is scored independently using a 0–4 scale based on two criteria: (A) whether the patented technology is hard to replicate, and (B) whether the technology is core to the business idea.

Marketing Group Protocol:

1. **Specificity** is first scored (0–4). If the score is ≤ 2, the evaluation ends here.

2. If Specificity ≥ 3, **Need Validity** is evaluated separately from two perspectives:

   - **ToC (Consumer)** needs: scored 0–3 based on the severity and importance of the need.
   - **ToB (Business)** needs: scored 0–3 based on the qualitative and quantitative return expected from addressing the need.

3. If Need Validity scores are too low (e.g., ToC = 1 or ToC + ToB ≤ 2), annotation ends. Otherwise, the annotator proceeds to **Market Size**, also evaluated from both ToC and ToB perspectives on a 0–3 scale.

Other Guidelines:

- Annotators were permitted to use external resources (e.g., web search, ChatGPT) to aid in evaluating technical feasibility or market need.

- Annotations focused on idea content, not linguistic quality. Minor grammatical issues or translation artifacts were not penalized.

- If a submission was truncated due to character limits and became incomprehensible, a low Specificity score (1 or 2) was assigned.

This pipeline-based protocol allowed evaluators to efficiently filter out infeasible ideas while focusing attention on higher-quality candidates.

Statistics of Human Evaluators

- **NLP / Computer Science:**
  - Technical group: 5 annotators (task organizers)
  - Marketing group: 7 annotators (consultants from Stockmark Inc.)

- **Material Chemistry:**
  - Combined group of 4 domain experts (Asahi Kasei Corporation)

2.3.3 LLM-as-a-Judge Evaluation

Models Used. To perform automated evaluation without relying on commercial APIs, we employed three open-access LLMs:

- `google/gemma-3-27b-it`

- `Qwen/Qwen3-30B-A3B`

- `meta-llama/Llama-3.3-70B-Instruct`

Inference Protocol. Each model was run with five different random seeds to ensure robustness. Two types of instructions were used:

- **Instruction #1 (Pairwise)**: Designed for direct comparison of two ideas. For each pair, two prompts were created by reversing the order of the ideas to mitigate positional bias. Inference results across seeds and orderings were aggregated via majority voting.

- **Instruction #2 and #3 (Scoring)**: Designed to assign a numerical score to each idea for specific evaluation criteria. The final score was computed as the mean across five sampled outputs.

This combination of human and automatic evaluation offers a reliable and scalable framework for assessing both the technical soundness and the business viability of LLM-generated product ideas.

3 Results and Discussion

We report the results of both automatic and human evaluations across the three domains—**NLP**, **Computer Science**, and **Material Chemistry**—with six evaluation criteria: *Technical Validity*, *Innovativeness*, *Specificity*, *Need Validity*, *Market Size*, and *Competitive Advantage*.

Domain	Criterion	1st	2nd	3rd
NLP	Tech. Validity	MK2 (1093)	MCG_DSN_late (1053)	ditlab (1010)
	Innovativeness	MK2 (1215)	ditlab (1111)	Shiramatsulab (1108)
	Specificity	MK2 (1215)	ditlab (1150)	Shiramatsulab (1113)
	Need Validity	MK2 (1076)	ditlab (1060)	Shiramatsulab (1030)
	Market Size	ditlab (1056)	TrustAI (1025)	MK2 (1008)
	Comp. Advantage	MK2 (1150)	MCG_DSN_late (1075)	ditlab (1034)
Computer Science	Tech. Validity	MK2 (1107)	ditlab (1003)	Shiramatsulab (983)
	Innovativeness	MK2 (1169)	ditlab (1078)	Shiramatsulab (1055)
	Specificity	MK2 (1170)	ditlab (1082)	Shiramatsulab (1007)
	Need Validity	MK2 (1053)	ditlab (1031)	Shiramatsulab (998)
	Market Size	TrustAI (1035)	MK2 (999)	ditlab (965)
	Comp. Advantage	MK2 (1124)	Shiramatsulab (1019)	ditlab (1011)
Material Chemistry	Tech. Validity	MK2 (1132)	ditlab (1021)	Shiramatsulab (998)
	Innovativeness	MK2 (1207)	MCG_DSN (1185)	NS_NLP (1152)
	Specificity	MK2 (1184)	MCG_DSN (1112)	ditlab (1067)
	Need Validity	NS_NLP (1129)	MK2 (1125)	ditlab (1093)
	Market Size	MK2 (1118)	ditlab (1050)	Shiramatsulab (1024)
	Comp. Advantage	MK2 (1146)	NS_NLP (1055)	ditlab (1011)

Table 1: Top three teams in automatic evaluation for each domain and criterion. Scores in parentheses.

Domain	Criterion	1st	2nd	3rd
NLP	Tech. Validity	MK2 (1025)	TrustAI (991)	ditlab (990)
	Innovativeness	MK2 (1103)	ditlab (1025)	TrustAI (926)
	Specificity	MK2 (1044)	ditlab (1036)	TrustAI (962)
	Need Validity	MK2 (1009)	ditlab (1003)	TrustAI (993)
	Market Size	TrustAI (1048)	ditlab (1024)	MK2 (921)
	Comp. Advantage	MK2 (1035)	ditlab (1008)	TrustAI (1000)
Computer Science	Tech. Validity	MK2 (1018)	TrustAI (1008)	ditlab (973)
	Innovativeness	MK2 (1036)	ditlab (992)	TrustAI (971)
	Specificity	ditlab (1020)	MK2 (995)	TrustAI (983)
	Need Validity	MK2 (1074)	TrustAI (980)	ditlab (945)
	Market Size	TrustAI (1035)	MK2 (999)	ditlab (965)
	Comp. Advantage	MK2 (1017)	ditlab (1007)	TrustAI (974)
Material Chemistry	Tech. Validity	TrustAI (1057)	MK2 (1017)	NS_NLP (1000)
	Innovativeness	NS_NLP (1017)	MCG_DSN (1009)	ditlab (1002)
	Specificity	ditlab (1047)	NS_NLP (1017)	MK2 (1010)
	Need Validity	ditlab (1035)	MCG_DSN (1026)	NS_NLP (1007)
	Market Size	NS_NLP (1017)	MK2 (1013)	ditlab (1009)
	Comp. Advantage	ditlab (1038)	TrustAI (998)	NS_NLP (997)

Table 2: Top three teams in human evaluation for each domain and criterion. Scores in parentheses.

3.1 Automatic Evaluation Results

Table 1 shows the top three systems for each domain and criterion in automatic evaluation. Across all domains, the **MK2** team achieved the highest average scores in nearly all criteria. Notably, in the NLP and Computer Science domains, MK2 consistently outperformed other teams, indicating a strong ability to generate ideas that aligned with LLM-based evaluation.

3.2 Human Evaluation Results

Table 2 summarizes the human evaluation results. In contrast to automatic evaluation, the rankings are more varied across domains, especially in Material Chemistry, where MK2 did not dominate.

Criterion	NLP	CS	Mat. Chem.
Tech Valid	-0.500	0.780	0.191
Innov	0.103	1.000	0.459
Spec	0.185	-0.155	0.049
Market Size	*	0.000	0.281
Need Valid	*	0.000	0.099
Comp Adv	0.563	-0.800	-0.199

Table 3: Krippendorff's α coefficientss in human evaluation for each domain and criterion.

3.3 Discussion

LLM vs. Human Judgment. In the NLP and Computer Science tracks, LLM-based and human evaluations aligned well, with MK2 and ditlab dominating both. However, in Material Chemistry, the human evaluators favored TrustAI and ditlab in criteria such as *Technical Validity* and *Competitive Advantage*, revealing a domain-specific gap in LLM judgment.

Inter-annotator agreement Table 3 shows Krippendorff's α coefficients (Krippendorff, 2011) for each evaluation criterion and domain, quantifying the consistency of human judgments. Overall, the coefficients are low With the exceptions of *Technical Validity* in Computer Science ($\alpha = 0.780$) and *Innovativeness* in Computer Science ($\alpha = 1.000$). These results indicate the subjectivity and difficulty of the remaining assessments.

Domain Expertise Matters. The Material Chemistry domain required deeper domain knowledge, which human annotators brought to bear. This underscores the limitations of general-purpose LLMs in specialized fields and motivates future work in domain-adapted LLMs or hybrid evaluation pipelines.

Specificity Drives Validity. Ideas with higher specificity tended to receive better evaluations across most other criteria. This confirms that concrete, well-described ideas are easier to evaluate and more likely to be perceived as feasible and valuable.

Takeaways for System Design. The strongest submissions incorporated structured prompts, external patent knowledge, and attention to both technical and business feasibility. Future systems may benefit from iterative generation, agentic collaboration, and retrieval-augmented generation with real-world context.

4 Conclusions

This paper presented an overview and analysis of the PBIG shared task, which challenges systems to generate realistic and creative product business ideas from patent data. The task provides a novel benchmark that spans technical feasibility, market reasoning, and creative synthesis—dimensions that are critical for real-world innovation.

Our analysis of the results revealed that while current LLMs can produce promising outputs, achieving balanced performance across technical and commercial criteria remains challenging. Top-performing systems like MK2 showcased how structured prompting, external knowledge use, and attention to business context can lead to strong results.

From an evaluation standpoint, combining human expert scores with LLM-based inference enabled scalable and fine-grained assessments, though subjectivity in some criteria remains a limiting factor. Continued research into robust, interpretable, and automated evaluation strategies is needed.

Looking ahead, we suggest the following directions for future research and shared tasks:

- Incorporating interactive or iterative ideation frameworks (e.g., multi-agent discussion, critique-and-revise).
- Using retrieval-augmented generation with market or user data for grounding.
- Enhancing the reproducibility and transparency of evaluation metrics.

We hope this shared task fosters further exploration into how language models can support the journey from technical invention to product innovation.

Acknowledgements

We thank the annotators for their valuable contributions to the data annotation process, which were essential for the evaluation of this shared task. From the Asahi Kasei TENAC R&D Department, we gratefully acknowledge Yoshitaka Shigematsu, Motohiro Fukai, Tomohito Asakawa, and Shinya Akebi. From Stockmark Inc., we thank Daisuke Nakagawa, Jiro Kashiwagi, Kumiko Boshita, Shotaro Kusaura, Tomoaki Okamoto, Wataru Inoue, and Yutaka Mori for their support and insights.

References

Tatsuya Ishigaki, Suzuko Nishino, Sohei Washino, Hiroki Igarashi, Yukari Nagai, Yuichi Washida, and Akihiko Murai. 2022. Automating horizon scanning in future studies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 319–327, Marseille, France. European Language Resources Association.

Ueda Keisuke, Hirota Wataru, Asakura Takuto, Omi Takahiro, Takahashi Kosuke, Arima Kosuke, and Ishigaki Tatsuya. 2025. Exploring design of multi-agent llm dialogues for research ideation. In *Proceedings of SIGDIAL*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *ICLR*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

A Annotation Guidelines (Technical Evaluation)

Thank you for participating in the annotation for Product Business Idea Generation from patent documents (PBIG). This document explains how to evaluate ideas from a technical perspective.

1. Evaluation Flow

Ideas are to be evaluated in the following order. The scoring criteria are defined in Section 2.

1. Evaluate **Specificity**.
    - If the score is 0, 1, or 2, stop the evaluation (do not proceed to Technical Validity or Innovativeness).
    - If the score is 3 or 4, proceed to Technical Validity.
2. Evaluate **Technical Validity**.
    - If the score is 0 or 1, stop the evaluation (do not proceed to Innovativeness).
    - If the score is 2 or higher, proceed to Innovativeness.
3. Evaluate **Innovativeness**.
4. Evaluate **Competitive Advantage** independently.

2. Scoring Definitions

Specificity (0–4)

- 0: Cannot judge / Insufficient background knowledge.
- 1: Not readable as natural language.
- 2: Readable, but intention is unclear and no concrete product is imaginable.
  Example: "A platform contributing to carbon neutrality."
- 3: One or more specific product ideas can be imagined (some ambiguity remains).
  Example: "A tool for obtaining user insight from social media."
- 4: One clearly defined product is imagined.
  Example: "A washing machine operable by voice commands."

Technical Validity (0–4)

- 0: Cannot judge.
- 1: Patent technology does not appear applicable or is irrelevant.
- 2: Difficult to implement, but prototyping is feasible.
- 3: Prototyping is feasible using the patented technology.
- 4: Production-ready implementation is feasible.

Innovativeness (0–5)

- 0: Cannot judge.
- 1: Already known application; lacks novelty.
- 2: Known applications exist but underexplored.
- 3: Unusual use case, but not especially novel.
- 4: Interesting and surprising idea.
- 5: Highly innovative idea.

Competitive Advantage (0–4) Evaluate based on the following two criteria:

Criterion A: Is it difficult to replicate the business idea without the patented technology?

Example (Fails A): Extracting date mentions from text – easily replaceable by general-purpose NLP tools.

Example (Satisfies A): Making accurate recommendations with few labeled samples – difficult to replicate.

Criterion B: Is the patented technology essential to realizing the business idea?

Example (Fails B): Reducing speaker weight to improve car fuel efficiency – the component contributes minimally.

Example (Satisfies B): Reducing main body weight to improve car fuel efficiency – major impact on outcome.

Then assign a score according to the combination:

- 0: Cannot judge.

- 1: Neither A nor B is satisfied.
- 2: Only B is satisfied (the technology is core, but not strong).
- 3: Only A is satisfied (the technology is strong, but not core).
- 4: Both A and B are satisfied.

B Annotation Guidelines (Marketing Evaluation)

Thank you for participating in the annotation for PBIG. This document explains how to evaluate ideas from a market perspective.

1. Column Definitions

- `idea_id`, `patent_number`, `patent_title`, `patent_abstract`: Not used in market evaluation.
- `idea_title`, `idea_description`, `idea_implementation`, `idea_differentiations`: These are the elements being evaluated.

2. Evaluation Flow

1. Evaluate **Specificity**.
   If Specificity ≤ 2, stop here.
2. Evaluate **Need Validity** (ToC and ToB perspectives).
   If ToC = 1 or ToC + ToB ≤ 2, stop here.
3. Evaluate **Market Size** (ToC and ToB perspectives).

3. Scoring Definitions

Specificity (0–4) Same as in technical evaluation.

Need Validity (ToC)

- 0: Cannot judge / no ToC relevance.
- 1 (Low): Weak need, few seek solutions.
  Example: "Earphone cables tangle sometimes."
- 2 (Medium): Some burden, but not critical.
  Example: "Shoulder pain from computer use."
- 3 (High): Severe or essential need.
  Example: "Fall risk for elderly at home."

Need Validity (ToB)

- 0: Not a ToB idea / Cannot judge.
- 1 (Low): Minimal qualitative/quantitative benefit.
- 2 (Medium): Either qualitative or quantitative return is large.
  Examples: "Cost savings" (quantitative), "Knowledge transfer" (qualitative).
- 3 (High): Both types of return are large.

Market Size (ToC)

- 0: Cannot judge / not a ToC product.
- 1 (Small): Niche or non-essential item.
  Example: "VR goggles, road bikes"
- 2 (Medium): Popular, not essential.
  Example: "Tablets, coffee makers"
- 3 (Large): Nearly all households need it.
  Example: "Toothbrushes, smartphones"

Market Size (ToB)

- 0: Not a ToB product / Cannot judge.
- 1 (Small): Useful to a few companies.
  Example: "Fast PoC for car parts"
- 2 (Medium): Addressable need for many, but conditional.
  Example: "BI tools"
- 3 (Large): Needed by most companies.
  Example: "Procurement management tools"

4. Annotation Notes

- Annotators may use Google or ChatGPT to aid judgment.
- If technical content is unclear, market evaluation should still proceed.
- Do not penalize for minor unnatural Japanese or truncation artifacts.
- If truncation makes the idea meaningless, assign Specificity = 1 or 2.

5. Example Cases

- **Case 1:** Specificity = 2 $\Rightarrow$ stop.
- **Case 2:** Specificity = 3, ToC + ToB = 2 $\Rightarrow$ continue to Market Size.