Collaborative Invention: Team ditlab at the AgentScen Shared Task – Refining Patent-based Product Ideation via LLM-Guided Selection and Rewriting

Yasunori Terao
DENSO IT Laboratory
Tokyo, Japan
terao.yasunori@core.d-itlab.co.jp

Yuuki Tachioka
DENSO IT Laboratory
Tokyo, Japan
tachioka.yuki@core.d-itlab.co.jp

Abstract

Our team, ditlab, participated in the AgentScen Shared Task. We propose a two-stage system for generating product ideas from patents, developed for the PBIG task. Patent texts pose challenges due to their technical density and limited focus on user value. Our method addresses this by combining diverse idea generation and pairwise comparison by large language models (LLMs) with guided refinement using a different type of LLM. Experimental results show strong performance, especially in specificity and innovation, and demonstrate that refinement with heterogeneous LLMs is effective in improving the quality of ideas. These findings highlight the potential of collaborative multi-LLM workflows for structured ideation from complex technical documents.

1 Introduction

Generating product ideas grounded in existing patents is a promising yet challenging task. Patents are rich sources of technical insight, but transforming this technical content into viable business ideas requires a combination of domain understanding, creativity, and user-centric thinking. Although recent advances in large language models (LLMs) have shown success in scientific discovery and ideation tasks (Si et al., 2024; Wang et al., 2024), the generation of product business ideas from patents remains relatively underexplored.

We participated in the Product Business Idea Generation from Patents (PBIG) (Chen et al., 2025) task at the AgentScen workshop as "ditlab" team. In this task, a system receives a patent document as input and outputs four concise descriptions corresponding to a product name, its function and target users, an implementation plan, and a point of differentiation from existing solutions. These outputs are evaluated by both humans and LLMs using multiple criteria. Since each field is subject to strict character limits, incorporating diverse evaluation aspects in a compact and effective manner poses a unique challenge.

To address this, we propose a method for generating and refining product ideas with the collaboration of multiple LLMs. The workflow can be divided into two steps: candidate generation and idea refinement. In the first stage, we generate diverse candidate ideas using different prompting strategies and LLMs and evaluate these ideas through pairwise comparisons with a strong baseline, using LLM-based judgments to identify higher-quality outputs. In the second stage, we independently generate auxiliary ideas using a different type of LLM and use them as references to guide further refinement. We re-evaluate the refined ideas and select the final output based on quality scores, ensuring that only improvements over the baseline are retained.

This framework is designed to systematically select and polish promising ideas, balancing multiple evaluation dimensions while adhering to the strict format constraints of the PBIG task. In the following sections, we detail our system design, evaluation process, and observations.

2 System Overview

Our system is designed as a two-stage pipeline that integrates idea generation, pairwise evaluation, preliminary selection of high-quality ideas, and refinement using multiple LLMs.

2.1 Candidate Generation and Evaluation (First Stage)

Figure 1 illustrates the workflow of the first stage: candidate generation and evaluation. We begin by generating product ideas from each patent using four prompting configurations:

1. **GPT-4.1 with the baseline prompt (baseline)**: Only the `description` field of the patent is provided as input.

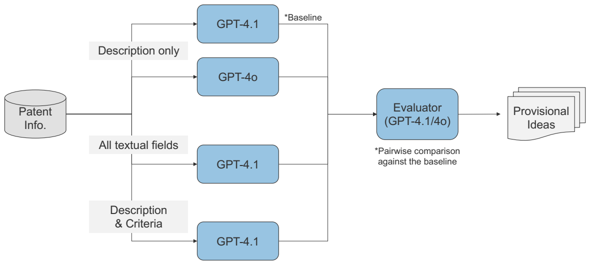

Figure 1: Candidate generation and evaluation stage (first stage). GPT-4.1 and GPT-4o generate four ideas per patent with different prompting strategies. After the evaluation through pairwise comparisons against a baseline, the best performing idea is selected as the provisional ideas.

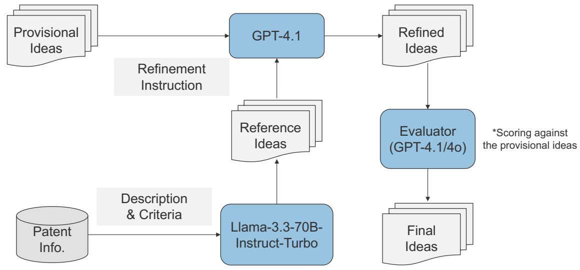

Figure 2: Refinement and final selection stage (second stage). Each provisional idea is refined using GPT-4.1 with reference to an independently generated idea from Llama-3.3-70B-Instruct-Turbo. Both are scored by GPT-4.1 and GPT-4o, and the highest-rated idea is selected as the final output.

2. **GPT-4o with the same prompt**: Identical to the baseline setting, but GPT-4o is used.

3. **GPT-4.1 with full patent text**: All textual content of the patent is provided, excluding any images. This design aims to capture broader contextual information while conserving token usage. We believe that image content has limited added value under strict character constraints.

4. **GPT-4.1 with evaluation criteria**: The baseline prompt is extended to include a brief explanation of the official evaluation criteria, encouraging the model to optimize the outputs accordingly.

All models were accessed through the OpenAI API.

Each idea generated under the above settings is evaluated in a pairwise comparison against the baseline output, using GPT-4.1 and GPT-4o as judges. The evaluation prompt is a lightly modified version of the official example provided by the organizers. For each comparison, LLM judges which is better (win or loss) or both are comparable (tie).

Ideas that outperform the baseline are selected as provisional ideas. If multiple such ideas exist for a given patent, one is randomly chosen to represent the best-performing candidate at this stage. If none of the ideas generated beat the baseline, the baseline itself is retained as the provisional idea.

2.2 Refinement and Final Selection (Second Stage)

Figure 2 illustrates the workflow of the second stage: refinement and final selection. To mitigate the potential bias arising from relying solely on ChatGPT-based models, we introduce an additional round of idea generation using Llama-3.3-70B-Instruct-Turbo (Grattafiori et al., 2024). This model is prompted by the same prompt with evaluation criteria (4), which is the most effective in the experiment in Table 2, but generates ideas independently from the previous stages.

For each patent, the selected provisional idea and the Llama-generated idea are both embedded into a refinement prompt and passed to GPT-4.1. GPT-4.1 is instructed to improve the provisional idea with reference to the Llama output, if such an improvement appears warranted, particularly in terms of fluency, specificity, and alignment with user needs.

We then evaluate both the refined idea and the original provisional idea using GPT-4.1 and GPT-4o, assigning quality scores on a 5-point scale (1 to 5) in increments of 0.1. This scoring-based evaluation replaces the earlier win-loss-tie format, which often resulted in ties that were difficult to resolve.

If both GPT-4.1 and GPT-4o assign higher scores to the provisional idea, it is retained as the final output, otherwise, the refined idea is selected. The result of this filtering constitutes the final idea submitted for each patent.

3 Experiments

3.1 Experimental Setups

The participants were given 150 patents from the United States Patent and Trademark Office, evenly drawn from three technical domains: materials chemistry (matchem), natural language processing (nlp), and computer science (cs). Each patent included full textual content and associated figures. The task was to generate one plausible product idea per patent that could realistically be launched within three years. The required output included a product title (up to 100 characters), a product description summarizing key features, target users, needs, and benefits (up to 300 characters), an implementation description detailing how the patented technology would be applied (up to 300 characters), and a differentiation statement explaining the uniqueness of the solution (up to 300 characters).

The evaluation was carried out by scoring ideas by human experts from the technical and market group and non-commercial LLM [1]. The evaluation criteria were technical validity, innovativeness, specificity, need validity, market size, and competitive advantage. The final rankings were calculated using Elo scoring based on judgments such as "Idea A is better," "Idea B is better," "Tie," or "Neither is good."

3.2 Official Evaluation Results

This section provides the official evaluation results and a brief summary of our observations. Table 1 summarizes the results of our system as extracted from the official leaderboard, across all categories and evaluation criteria. The following are key observations based on these results. In general, Elo scores are relatively higher in automatic evaluation (`auto-*`) than human evaluation, indicating that our system aligned well with LLM-based evaluators. In `auto-nlp`, the system achieved a particularly strong performance in *specificity* (1150) and *innovation* (1111), suggesting an effective generation of concrete and novel ideas.

In human evaluation categories (`human-*`), the scores exhibit more variability, possibly due to subjective differences among the annotators. In the `human-matchem` category, our system ranked first in *specificity*, *need validity*, and *competitive advantage*, indicating that in terms of some aspects our results were positively received by human judges. In the `human-cs` category, the system received relatively low scores in *need validity* (945) and *market size* (965), suggesting room for improvement in articulating user needs and market feasibility.

3.3 First Stage Results

When comparing the ideas generated using the same prompt for GPT-4.1 and GPT-4o, we found that both GPT-4.1 and GPT-4o tended to rate the ideas generated by GPT-4.1 as superior. This suggests a consistent preference for GPT-4.1 ideas between both evaluators.

The titles of the ideas generated by LLMs frequently included the suffix "Pro," such as in "VisionFit Pro" or "DataSpeak Pro". This naming pattern was consistently observed across different patent domains, suggesting a broad tendency toward professional-sounding or premium-style branding.

Table 2 shows the counts of the win/loss/tie results across domains and configurations. GPT-4.1 was used to evaluate LLM-generated ideas. GPT-4o with the baseline prompt (Prompt 2) performed worse than the baseline (Prompt 1), especially in the CS and NLP domains. In contrast, Prompt 4, which includes explicit evaluation criteria, showed moderate improvements.

Table 3 shows the counts of the win/loss/tie results evaluated by GPT-4o. The tendencies were similar to those in Table 2 but the differences between the evaluators were observed: GPT-4.1 tended to favor its own refinements, while GPT-4o considered the ideas generated by Prompt (2) more preferable for some cases. It has been known that LLM judges tend to favor the answers generated by themselves. These evaluator preferences should be considered when using LLMs as judges (Ye et al., 2024).

For both tables, the refinement of ideas in corporation with Llama-3.3 was the most effective strategy, achieving the highest counts of win across all domains. This suggests that incorporating perspectives from heterogeneous LLM can enhance the diversity of generated ideas rather than increasing the diversity of prompts for the same type of LLM.

3.4 Second Stage Result

Table 4 shows the counts of win/loss/tie. In most cases, the refined ideas were evaluated better than their provisional counterparts in the first stage, which shows the effectiveness of our proposed refinement using a different type of LLM.

4 Conclusion and Future Work

We proposed a two-stage framework for patent-based product ideation that combines diverse LLM generation with guided refinement using auxiliary models. Our system performed well in automatic evaluations, particularly in specificity and innova-

[1]Open source LLMs were used: https://huggingface.co/google/gemma-3-27b-it https://huggingface.co/Qwen/Qwen3-30B-A3B, and https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

Category (n)	tech_valid	spec	neeed_valid	market_size	innov	comp_adv
auto-matchem (7)	1021 (3)	1067 (3)	1093 (3)	1050 (2)	1052 (4)	1011 (3)
auto-nlp (7)	1010 (4)	1150 (2)	1060 (3)	1056 (2)	1111 (2)	1034 (3)
auto-cs (5)	1003 (2)	1082 (2)	1031 (2)	1015 (3)	1078 (2)	1011 (3)
human-matchem (5)	996 (4)	1047 (1)	1035 (1)	1009 (3)	1002 (3)	1038 (1)
human-nlp (4)	990 (4)	1036 (2)	1003 (2)	1024 (2)	1025 (2)	1008 (2)
human-cs (3)	973 (3)	1020 (1)	945 (3)	965 (3)	992 (2)	1007 (2)

Table 1: Official Elo-based evaluation scores for each patent domain and evaluation type (automatic or human). The columns correspond to the evaluation criteria: technical validity, specificity, need validity, market size, innovativeness, and competitive advantage. Each cell shows the Elo score, and parentheses indicate the number of participating systems and the rank within the category.

Config	matchem	nlp	cs
Prompt (2)	3 / 46 / 1	1 / 48 / 1	0 / 48 / 2
Prompt (3)	15 / 22 / 13	7 / 7 / 36	9 / 8 / 33
Prompt (4)	20 / 5 / 25	12 / 7 / 31	13 / 8 / 29
Refinement	41 / 0 / 9	37 / 1 / 12	35 / 0 / 15

Table 2: The counts of win/loss/tie by domain for each configuration compared with the ideas generated with prompt (1) with GPT-4.1. The evaluation was carried out using GPT-4.1. Prompts (2), (3), and (4) correspond to those in Section 2.1.

Config	matchem	nlp	cs
Prompt (2)	1 / 37 / 12	0 / 44 / 6	0 / 44 / 6
Prompt (4)	3 / 5 / 42	3 / 10 / 37	7 / 12 / 31
Refinement	19 / 3 / 28	11 / 3 / 36	13 / 2 / 35

Table 3: The configurations are the same as those in Table 2 but evaluation was carried out by GPT-4o. The evaluation of Prompt (3) was omitted in this case.

Eval. by	matchem	nlp	cs
GPT-4.1	43 / 7 / 0	43 / 7 / 0	47 / 3 / 0
GPT-4o	41 / 9 / 0	33 / 17 / 0	43 / 7 / 0

Table 4: The counts of win/loss/tie by domain for the refined ideas using Llamma-3.3 compared with the provisional ideas in the first stage.

tion, and was highly ranked in human evaluations for the material chemistry domain.

Refinement through a different LLM consistently improved idea quality, highlighting the value of cross-model collaboration over prompt variation alone. We also observed naming trends and evaluator biases that favor the generation model, suggesting the need for evaluator diversification.

Future work will explore role-specialized LLMs for generation, critique, and refinement to improve output quality and mitigate model-specific biases.

References

Chung-Chi Chen, Tatsuya Ishigaki, Sophia Ananiadou, and Hiroya Takamura. 2025. Product business idea generation from patents (PBIG). `https://sites.google.com/view/agentscen/shared-task`. Accessed: 2025-06-20.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. *Preprint*, arXiv:2409.04109.

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)*, page 279–299. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *Preprint*, arXiv:2410.02736.

A Score Results in Second Stage

Fig. 3 shows the rated scores for all patents in each category. The scores of almost all ideas ranged between 4.3 and 4.8, although we requested the LLM to evaluate the ideas on a scale of 1-5. If this score is precise, the improvements were generally modest.

Table 5 shows the average and standard deviation of scores in all domains. In all cases, the scores for the refined ideas are better than those for the baseline ideas. In general, the scores evaluated by GPT-4o are consistently lower than those of GPT-4.1, and the differences between the baseline and refined versions are also smaller under GPT-4o. Scores are similar in the NLP and CS domains. In contrast, the matchem domain shows distinctly lower scores, possibly due to its higher level of domain expertise required. However, refined ideas in the matchem domain still received relatively high scores under GPT-4o, indicating the effectiveness of refinement even in specialized fields.

B Examples of Generated Ideas

Figures 4, 5, and 6 show examples of idea refinement with the greatest score improvements in matchem, nlp, and cs, respectively. The scores of these ideas, evaluated by GPT-4.1, improved by 0.4 to 0.5 points through the refinement process. Across the three themes analyzed, it is evident that the Refined Ideas (C) consistently build upon the Provisional Ideas (A), enhancing their practicality and scalability. Although the ideas generated by Llama-3.3 (B) do not appear to directly influence the refined versions, they do play a meaningful role in stimulating broader thinking and offering alternative perspectives for extending or generalizing the initial concepts.

One clear pattern across all Refined Ideas is the use of the suffix "Pro" in their product names. This consistency is unlikely to be coincidental. Rather, it reflects GPT-4.1's implicit preference for positioning the refined outputs as high-performance, commercially viable versions of the original ideas.

Idea	matchem	nlp	cs
Eval. by GPT-4.1			
Baseline	4.48 ± 0.16	4.58 ± 0.10	4.57 ± 0.12
Refined	4.67 ± 0.11	4.73 ± 0.08	4.74 ± 0.08
Eval. by GPT-4o			
Baseline	4.39 ± 0.14	4.43 ± 0.15	4.42 ± 0.14
Refined	4.55 ± 0.13	4.52 ± 0.10	4.57 ± 0.12

Table 5: The average and standard deviation of scores in terms of domain and evaluators.

Examining the content of the Refined Ideas reveals a distinctive structural shift. While the Provisional Ideas tend to focus on the technology and its immediate use case, the Refined versions expand on this by explicitly addressing who the users are, in what contexts the products are deployed, and what real-world problems they solve. For example, in the NLP case, although the core classification and explanation functions remain the same, the refined version targets the enterprise segment and incorporates terms such as audit readiness, compliance, and enterprise-scale deployment, defining the tool within the context of organizational trust and accountability. Similarly, in the case of matchem, the refined idea incorporates terms such as Industry 4.0 to align the product with broader industrial trends and visions.

In contrast, the B ideas tend to be more abstract and less grounded in specific commercial or operational use cases. For example, the target of the NLP B idea, "scientists and researchers", is a narrower and less commercially attractive market. In the cs example, Idea B refers broadly to "electronic devices" without specifying which industries or applications would benefit the most. Although this vagueness may be a limitation in terms of market clarity, it also provides conceptual flexibility, allowing new use cases and variations to emerge. Notably, elements such as the template-based NLG from the NLP B idea or the variable-depth compressive stress layer from the glass cover B idea may not appear explicitly in the Refined Ideas, but likely serve as conceptual input that enriches the refinement process.

Comparing the three case studies highlights how the Refined Ideas evolve the Provisional ones. In the cs example, A is focused narrowly on wearables, while C extends the scope to "critical medical sensors". C addresses failure modes such as breakage, leakage, and warping with a material engineering solution optimized for biomedical environments. The result is a concept that is both technically robust and clearly aligned with the unmet needs in the target domain.

Taken together, these findings suggest that Refined Ideas are polished versions of the Provisional Ideas and combine technical validity with deployment readiness and market relevance. The B ideas, while not directly mirrored in the refined ideas,

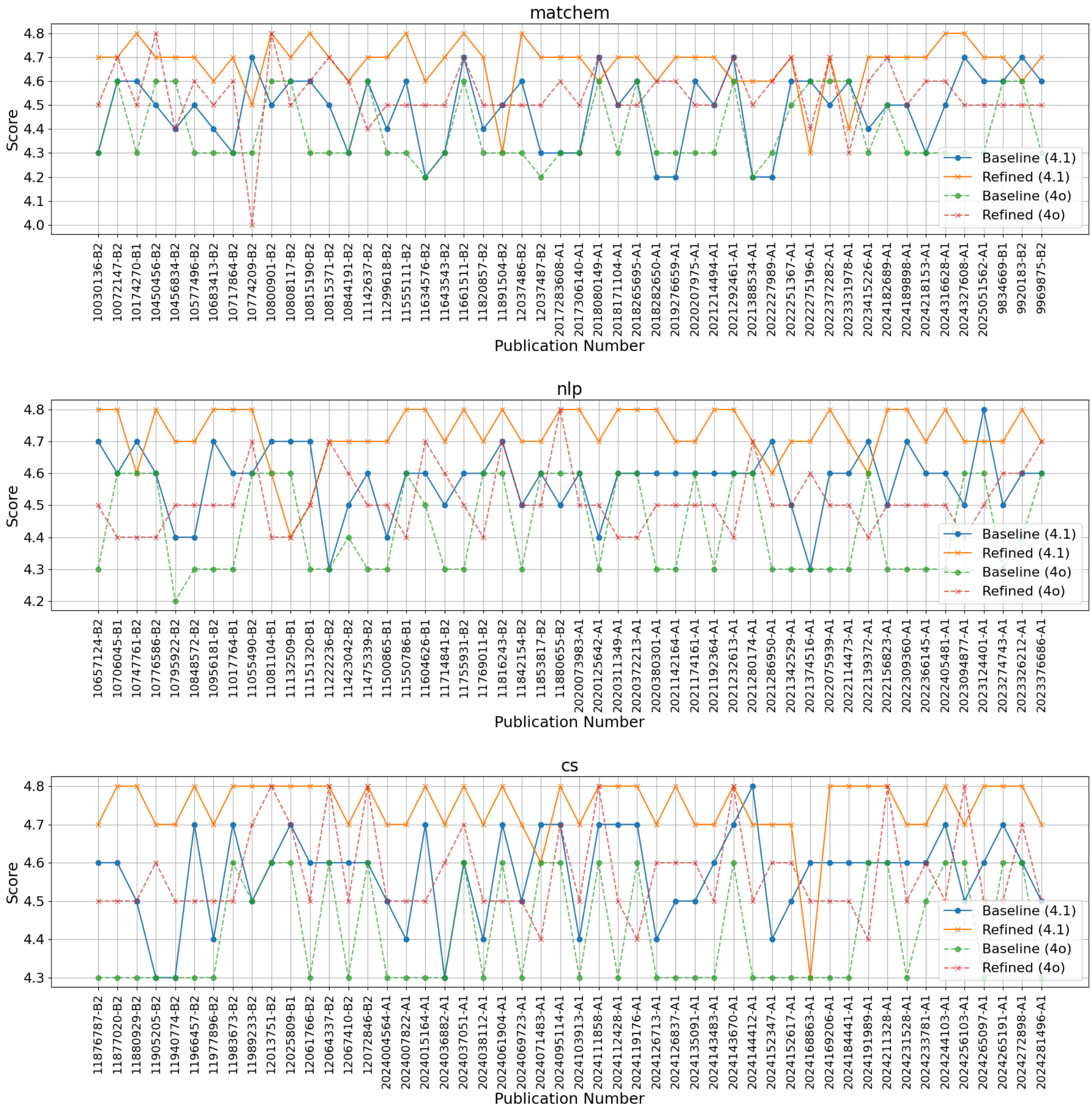

Figure 3: Scores of second stage for each category.

contribute by offering abstract concepts and broadening the thinking space.

C Prompts Used for Generation and Evaluation

The prompts for generation are shown in Figures 7–10, and those for evaluation in Figures 11 and 12.

(A) Provisional idea

Product Title: SnapGear: High-Performance Modular Gears for Collaborative Robotics
Product Description: Precision-molded polyacetal gears with superior impact resistance, thermal stability, and easy moldability for collaborative robot joints in factories. Targeted at robotics OEMs needing reliable, long-life, low-maintenance drive components.
Implementation: Utilize the patented polyacetal resin composition to injection mold gears with complex tooth geometries and integrated mounting features, enabling direct use in robot joints that require high impact strength, heat resistance, and smooth operation.
Differentiation: Unlike standard gears, SnapGear delivers higher durability, reduced yellowing, and superior mold release, enabling cost-effective mass production of complex, lightweight gears optimized for demanding collaborative robotics environments.

(B) Idea generated by Llama-3.3

Product Title: SmartGear
Product Description: A line of high-performance, impact-resistant gears for industrial machinery and automotive applications, utilizing the patented polyacetal resin composition for enhanced thermal stability and releasability.
Implementation: The patented polyacetal resin composition will be used to manufacture the gears, providing improved toughness and resistance to deformation. The unique blend of antioxidants, nitrogen-containing compounds, and fatty acid amides will ensure optimal performance in high-temperature environments.
Differentiation: SmartGear stands out from existing solutions due to its unparalleled combination of impact resistance, thermal stability, and moldability, making it an ideal choice for demanding industrial and automotive applications where reliability and efficiency are crucial.

(C) Refined idea

Product Title: SnapGear Pro: Self-Lubricating Precision Gears for Collaborative Robots & Smart Automation
Product Description: Injection-molded polyacetal gears with built-in lubrication and high impact/thermal resistance, designed for next-gen collaborative robots and automated machinery. Delivers quieter, longer-lasting, maintenance-light performance for OEMs and factories.
Implementation: Leverage the patented polyacetal resin composition, including fatty acid amide for internal lubrication, to mold gears with complex features. Optimize process for superior release, impact strength, and low friction—ideal for robot joints and conveyor drives.
Differentiation: Unlike standard or even advanced gears, SnapGear Pro combines intrinsic lubrication, high durability, and stable performance under heat, reducing downtime and maintenance—key for robotics and Industry 4.0 automation where reliability is paramount.

Figure 4: Idea refinement for 2019276659-A1 in matchem.

(A) Provisional idea

Product Title: ExplainAI: Transparent Document Classifier for Regulated Industries
Product Description: ExplainAI is an NLP-powered text classification tool for legal, healthcare, and finance professionals, providing automated document categorization with clear, natural language explanations and user feedback loops for improved trust and compliance.
Implementation: By integrating the patented micro/classifier system, ExplainAI assigns document labels and generates human-readable explanations for each decision. User feedback helps refine explanations and prunes ineffective micro classifiers, enhancing accuracy over time.
Differentiation: Unlike black-box NLP tools, ExplainAI uniquely delivers transparent, user-validated rationale for every classification, supporting regulatory compliance, audit trails, and user trust in high-stakes industries where explainability is critical.

(B) Idea generated by Llama-3.3

Product Title: Explainify
Product Description: Explainify is an AI-powered tool that provides natural language explanations for statistical classifier predictions, helping data scientists and researchers understand and trust their models' decisions.
Implementation: Explainify will utilize the patented technology to train a statistical classifier and a set of micro classifiers on labeled and unlabeled data, generating explanations for predictions using a template-based natural language generation technique.
Differentiation: Explainify stands out from existing solutions by offering transparent and interpretable AI decisions, enabling users to identify biases and errors, and improving model performance over time through user feedback and continuous learning.

(C) Refined idea

Product Title: ExplainAI Pro: Trustworthy Document Intelligence for Compliance-Driven Enterprises
Product Description: ExplainAI Pro empowers legal, healthcare, and financial teams to auto-categorize documents with AI-generated, user-validated natural language explanations, ensuring regulatory compliance, audit readiness, and user trust in high-stakes workflows.
Implementation: Utilizing the patent's dual-classifier and micro-classifier approach, ExplainAI Pro delivers document labels alongside clear, tailored explanations. Integrated user feedback refines explanations and model accuracy, supporting large-scale enterprise deployments.
Differentiation: Unlike generic NLP tools, ExplainAI Pro uniquely combines explainable AI, user-driven feedback, and audit-ready trails, giving compliance-critical industries transparent, continuously improving classification for both regulatory and operational advantage.

Figure 5: Idea refinement for 2021374561-A1 in nlp.

(A) Provisional idea

```
Product Title: SealSure: Ultra-Flat, Chemically Strengthened Glass Covers for Medical Wearables
Product Description: SealSure offers custom-shaped, ultra-flat glass covers for medical wearables (e.g., glucose monitors, ECG patches) that ensure hermetic sealing, high impact resistance, and reliable sensor function for patients and clinicians.
Implementation: We apply the dual ion-exchange and local material removal process to manufacture glass covers with precisely flat mounting surfaces, tailored for medical wearables that require robust sealing and durability over extended wear.
Differentiation: Unlike generic covers, SealSure guarantees micron-level flatness for leak-proof seals crucial in medical devices, combining advanced impact resistance with biocompatibility and custom geometries, solving both breakage and sealing failures.
```

(B) Idea generated by Llama-3.3

```
Product Title: SmartCover
Product Description: A chemically strengthened glass cover for electronic devices, providing enhanced durability and resistance to warping, with a unique mounting surface design for seamless integration.
Implementation: Utilizing the patented ion-exchange technology to create a compressive stress layer, ensuring the cover's strength and flatness, while incorporating a specialized polishing process to maintain the mounting surface's planarity.
Differentiation: SmartCover stands out with its innovative material removal process, enabling the creation of a variable-depth compressive stress layer, resulting in unparalleled strength and durability, setting it apart from conventional glass covers.
```

(C) Refined idea

```
Product Title: SealSure Pro: Precision-Engineered Glass Covers for Critical Medical Sensors
Product Description: SealSure Pro delivers ultra-flat, chemically strengthened glass covers tailored for advanced medical sensors (e.g., continuous glucose monitors, cardiac patches), ensuring leak-proof seals, durability, and reliable sensor readings for healthcare providers and patients.
Implementation: We leverage the dual ion-exchange with region-specific material removal to achieve micron-level flatness and custom compressive stress profiles, optimizing sealing surfaces and mechanical strength for each device geometry and medical application.
Differentiation: SealSure Pro uniquely combines patent-driven variable-depth compressive stress layers with biocompatible designs to guarantee both impact resistance and hermetic, distortion-free sealing—solving failures that generic covers or coatings can't address in medical wearables.
```

Figure 6: Idea refinement for 11905205-B2 in cs.

```
I give you the description of a patent. Read it.

<patent>
description
</patent>

## Task

Generate one business idea for a product using this patent.

Output the idea in the following format:

{
"product_title": "...",
"product_description": "...",
"implementation": "...",
"differentiation": "..."
}

## Rules

- product_title: A concise name for your product (up to 100 characters).
- product_description: A brief explanation of the product outlining its
essential features and functions, the target users, their needs, and the
benefits provided by the product (up to 300 characters).
- implementation: An explanation describing how you will implement the
patent's technology into your product (up to 300 characters).
- differentiation: An explanation highlighting what makes your product
unique and the reason why it stands out from existing solutions (up to 300
characters).
```

Figure 7: Baseline prompt for idea generation.

```
I will give you structured information about a patent. Please read it
carefully and use it to generate one product idea.

## Patent
<patent_title>
{{ title }}
</patent_title>


<abstract> {{ abstract }}
</abstract>

<claims>
{{ claims }}
</claims>

<description>
{{ description }}
</description>

## Task
Use the patent information to propose a **new product idea** that applies
the patented technology.

- Use the **title and abstract** to understand the general scope of
the invention.
- Use the **claims** to understand what makes the invention unique or legally
protected.
- Use the **description** to understand possible implementations and
technical details.

## Output Format
Please output your idea in the following JSON format:

“‘json
{ "product_title": "...",
"product_description": "...",
"implementation": "...",
"differentiation": "..."
} “‘

## Rules
- product_title: A concise name for your product (up to 100 characters).
- product_description: A brief explanation of the product outlining its
essential features and functions, the target users, their needs, and the
benefits provided by the product (up to 300 characters).
- implementation: An explanation describing how you will implement the
patent’s technology into your product (up to 300 characters).
- differentiation: An explanation highlighting what makes your product
unique and the reason why it stands out from existing solutions (up to 300
characters).
```

Figure 8: Prompt with full patent text.

```
I give you the description of a patent. Read it.

<patent>
{{ description }} </patent>

## Task Generate one business idea for a product using this patent.
Your idea should not only be relevant to the technology described in the patent, but also designed with the following evaluation criteria in mind.

### Evaluation Criteria
1. **Technical validity** — Is the patent suitable for the product? Is the implementation feasible? Can it be done within three years?
2. **Innovativeness** — Does the patented technology offer a novel solution to the demand?
3. **Specificity** — Is the idea specific? For example, “help researchers manage references” is more specific than “help researchers do research.”
4. **Need validity** — Do the described users really need this solution?
5. **Market size** — Is the market large enough? Are there many potential users?
6. **Competitive advantage** — What business advantage does the product gain by using this patented technology?

## Output format
Present your idea in the following format:

“‘json
{
"product_title": "...",
"product_description": "...",
"implementation": "...",
"differentiation": "..."
}
“‘

### Rules
- product_title: A concise name for your product (up to 100 characters).
- product_description: A brief explanation of the product outlining its essential features and functions, the target users, their needs, and the benefits provided by the product (up to 300 characters).
- implementation: An explanation describing how you will implement the patent’s technology into your product (up to 300 characters).
- differentiation: An explanation highlighting what makes your product unique and the reason why it stands out from existing solutions (up to 300 characters).
```

Figure 9: Prompt with evaluation criteria.

```
I give you the description of a patent. Read it carefully.

<patent>
{{ description }}
</patent>

You have already proposed the following business idea based on this
patent:

<idea1>
{{ idea1 }}
</idea1>

Another person, based on the same patent, proposed this alternative
idea:

<idea2>
{{ idea2 }}
</idea2>

## Task
Your task is to **refine your own idea (idea1)** using the same patent and
evaluation criteria.
If the alternative idea (idea2) contains good elements that improve upon
your original idea — such as greater specificity, a stronger competitive
advantage, or better technical feasibility — you are encouraged to
incorporate them.
However, do not simply copy idea2. Instead, **use it as inspiration to
enhance your own idea**, while maintaining originality and grounding your
solution in the patent.

### Evaluation Criteria

...
```

Figure 10: Prompt for idea refinement.

```
## Inputs

Read (1) a patent and (2) two product business ideas using the technology in the patent.

<patent>
{{ patent.description }}
</patent>

<idea id=”1”>
{{ idea1 }}
</idea>

<idea id=”2”>
{{ idea2 }}
</idea>

## Task

Your task is to evaluate **both ideas across multiple criteria** and determine which one is better overall.

Please carefully consider the following evaluation criteria:

...


## Judgment Instructions

After reviewing all criteria, select **one overall judgment** based on the idea that performs better **across the board**.

Use the following judgment codes:

- ‘1‘: Idea 1 is better
- ‘2‘: Idea 2 is better
- ‘3‘: Tie (both are equally strong)
- ‘4‘: Neither is good (both ideas are weak)

## Output Format

Output your judgment in the following strict JSON format:

{
”judgement”: <1 or 2 or 3 or 4>,
”reason”: ”<reason explaining why you selected this judgment, ideally referencing multiple criteria>”
}
```

Figure 11: Prompt for evaluation (win/loss/tie).

```
## Inputs

Read (1) a patent and (2) two product business ideas using the technology in the patent.

<patent>
{{ patent.description }}
</patent>

<idea id="1">
{{ idea1 }}
</idea>

<idea id="2">
{{ idea2 }}
</idea>

## Task

Your task is to evaluate **both ideas across multiple criteria** and assign a total score to each idea using a 5.0-point scale (in increments of 0.1).

Please carefully consider the following evaluation criteria:

...


## Judgment Instructions

Assign a score between 0.0 and 5.0 (inclusive) to each idea, reflecting its overall quality across all six criteria.

- Use increments of 0.1 only (e.g., 4.5, 3.2, 0.7).
- Base your judgment on how well the idea satisfies the criteria as a whole.
- Then explain your reasoning for both scores, referencing specific criteria.

## Output Format

Output your judgment in the following strict JSON format:

{
"score_idea1": <float between 0.0 and 5.0>,
"score_idea2": <float between 0.0 and 5.0>,
"reason": "<reason explaining how each score was derived, referencing multiple criteria>"
}
```

Figure 12: Prompt for evaluation (score).