Aggregate vs. Personalized Judges in Business Idea Evaluation: Evidence from Expert Disagreement

Wataru Hirota[1], **Tomoki Taniguchi**[2], **Tomoko Ohkuma**[2], **Kosuke Takahashi**[1], **Takahiro Omi**[1], **Kosuke Arima**[1], **Takuto Asakura**[1], **Chung-Chi Chen**[3,4*], **Tatsuya Ishigaki**[4*]
[1]Stockmark Inc [2]Asahi Kasei Corporation [3] National Institute of Informatics
[4]National Institute of Advanced Industrial Science and Technology
wataru.hirota@stockmark.co.jp chen@nii.ac.jp ishigaki.tatsuya@aist.go.jp

Abstract

Evaluating LLM-generated business ideas is often harder to scale than generating them. Unlike standard NLP benchmarks, business idea evaluation relies on multi-dimensional criteria such as feasibility, novelty, differentiation, user need, and market size, and expert judgments often disagree. This paper studies a methodological question raised by such disagreement: should an automatic judge approximate an aggregate consensus, or model evaluators individually? We introduce PBIG-DATA, a dataset of approximately 3,000 individual scores across 300 patent-grounded product ideas, provided by domain experts on six business-oriented dimensions: specificity, technical validity, innovativeness, competitive advantage, need validity, and market size. Analyses show substantial expert disagreement on fine-grained ordinal scores, while agreement is higher under coarse selection, suggesting structured heterogeneity rather than random noise. We then compare three judge configurations: a rubric-only zero-shot judge, an aggregate judge conditioned on mixed evaluator histories, and a personalized judge conditioned on the target evaluator's scoring history. Across dimensions and model sizes, personalized judges align more closely with the corresponding evaluator than aggregate judges, and evaluator agreement correlates with similarity of judge-generated reasoning only under personalized conditioning. These results indicate that pooled labels can be a fragile target in pluralistic evaluation settings and motivate evaluator-conditioned judge designs for business idea assessment.

1 Introduction

Large language models (LLMs) make it easy to generate large numbers of product ideas. The practical bottleneck is not generation but evaluation. Organizations must decide which ideas are sufficiently concrete, technically feasible, differentiated, and commercially meaningful to justify further investment. These decisions are often made by multiple experts with different backgrounds, such as technical reviewers and business strategists.

Unlike tasks where evaluation targets factual correctness or task completion, business idea evaluation is multi-dimensional and judgment-driven. Even when reviewers follow the same rubric, they may apply different assumptions about feasibility, novelty, risk, or market opportunity. In practice, this leads to persistent disagreement across experts. Such disagreement is not necessarily annotation noise; it reflects heterogeneous standards about what constitutes a promising idea.

Most LLM-as-a-Judge approaches assume that a single scoring standard exists. Under this assumption, labels from multiple reviewers are aggregated, and the judge is optimized to reproduce this pooled signal. However, when expert disagreement is systematic, aggregation may obscure meaningful differences between evaluators. This raises a methodological question for evaluation design: should an automatic judge approximate an aggregate consensus, or should it model evaluators individually?

This paper studies this question in the context of business-oriented product idea evaluation. We introduce PBIG-DATA, a dataset of approximately 3,000 individual scores across 300 patent-grounded product ideas, provided by domain experts along six business-oriented dimensions: specificity, technical validity, innovativeness, competitive advantage, need validity, and market size.

Our empirical findings reveal a substantial gap between common evaluation assumptions and actual expert behavior. First, inter-annotator agreement on fine-grained ordinal scores is often close to zero and occasionally negative, indicating that expert scoring does not converge to a single shared scale. At the same time, agreement increases when the task is framed as coarse selection, suggesting that disagreement reflects heterogeneous but structured

*The last two authors are co-leads of this project.

standards rather than random noise. Second, this heterogeneity has direct implications for LLM-as-a-Judge design. We compare three judge configurations: a rubric-only zero-shot judge, an aggregate judge conditioned on mixed evaluator histories, and a personalized judge conditioned on the target evaluator's own scoring history. Across dimensions, personalized judges align more closely with the corresponding evaluator than aggregate judges. This indicates that individual evaluators are internally consistent even when they disagree with each other.

Taken together, these results suggest that business idea evaluation is inherently pluralistic. Treating pooled labels as a single ground truth may be an inadequate target for automatic judging. Modeling evaluator-specific standards provides a more faithful representation of expert judgment, but it also highlights how far current judge designs are from fully supporting heterogeneous decision processes in real-world settings.

This paper makes the following contributions:

1. A dataset of expert-scored product ideas under business-oriented criteria.
2. A quantitative analysis of structured expert disagreement.
3. Evidence that aggregate and personalized judge designs behave differently under heterogeneous standards, with implications for evaluation methodology in practical ideation systems.

2 Related Work

2.1 Human Evaluation in Creative Generation

Human evaluation plays a central role in creative natural language generation tasks. Prior surveys have shown that inter-annotator agreement is often low in creative settings, especially when tasks involve open-ended judgments such as novelty and emotional impact (Hämäläinen and Alnajjar, 2021). Amidei et al. (2019) argue that low agreement does not necessarily indicate unreliable data, but may reflect irreducible variability in human language interpretation. They recommend complementing agreement metrics with correlation-based analyses to better understand evaluation reliability.

Recent large-scale studies further confirm that even domain experts disagree substantially when evaluating generative outputs. For example, Si et al. (2025) report significant variation among NLP researchers assessing the novelty of LLM-generated research ideas. These findings suggest that disagreement is a structural property of creative evaluation rather than annotation noise. Our work extends this line of investigation to business-oriented product idea evaluation, focusing specifically on how disagreement affects the design of automatic judges.

2.2 LLMs for Ideation

LLMs have increasingly been used to support ideation tasks across domains. Interactive systems such as Wordcraft demonstrate how language models can assist human writers in creative writing workflows (Yuan et al., 2022). In scientific and research ideation, multi-agent LLM frameworks have been explored to improve diversity and feasibility of generated proposals (Ueda et al., 2025). Large-scale evaluations of idea generation benchmarks also examine LLM performance on novelty and feasibility metrics (Si et al., 2025).

While these studies focus primarily on improving generation quality, evaluation methodology remains comparatively underexplored. Most ideation benchmarks rely either on automatic scoring or on pooled human labels treated as ground truth. In contrast, our work centers on the evaluation stage itself and investigates how heterogeneous expert standards influence judge modeling.

2.3 LLM-as-a-Judge and Evaluation Modeling

LLMs are increasingly used as automatic judges for diverse NLP tasks. Early evidence shows that large models can approximate human preferences under certain prompting strategies (OpenAI, 2023). Subsequent studies systematically examine judge robustness and vulnerabilities, highlighting issues such as prompt sensitivity and alignment bias across domains (Thakur et al., 2025; Tan et al., 2025). Self-preference bias further complicates judge reliability when models evaluate outputs similar to their own generations (Wataoka et al., 2025).

Recent work explores whether LLMs can model individual evaluators rather than aggregate labels. Dong et al. (2024) investigate personalized judging and show that conditioning on evaluator-specific history can improve alignment. However, these studies primarily focus on general NLP evaluation tasks. The implications of personalization under systematically low inter-annotator agreement remain insufficiently studied.

Our work contributes to this discussion by examining aggregate versus personalized judge config-

urations in a business ideation setting. By explicitly quantifying expert disagreement and comparing judge alignment under heterogeneous standards, we provide evidence that evaluation modeling choices must account for pluralistic expert judgments rather than assuming a single unified ground truth.

3 Data and Evaluation Setup

This section describes the data, scoring dimensions, and annotation protocol used in our study. The objective is to define an evaluation setting where multiple experts score the same type of business idea under a shared rubric, while allowing for heterogeneous standards and incomplete consensus.

3.1 Scoring Dimensions and Rubric

Each idea is scored along six business-oriented dimensions: specificity, technical validity, innovativeness, competitive advantage, need validity, and market size, summarized in Table 1 (the full level-by-level rubric is provided in Appendix A). Our six dimensions are synthesized from two widely used frameworks for early-stage product and innovation assessment applied to LLM-generated patent-grounded ideas. The NABC framework (Carlson and Wilmot, 2006) frames new product assessment around Need, Approach, Benefit, and Competition, emphasizing that promising ideas should combine a concrete user need, a credible technical approach, differentiated benefits, and awareness of the competitive landscape. Cooper's Stage-Gate model (Cooper, 1990) operationalizes early-stage screening along feasibility, differentiation, and market attractiveness criteria.

The scale for each dimension is chosen to match its natural granularity in screening decisions rather than to enforce a uniform range. Specificity, technical validity, and competitive advantage use 1–4 scales because they admit natural four-level gradations (from "unusable" to "production-ready" in the case of technical validity, for example). Innovativeness uses a 1–5 scale to give evaluators an additional level to distinguish "surprising but not groundbreaking" from "clearly innovative". Need validity and market size use 0–3 scales because the lowest level encodes a categorical exclusion ("not a B2B product") rather than a low-quality gradation; collapsing this into a 1-based scale would conflate "non-applicable" with "applicable but weak".

Dimension	Scale	Threshold	Focus
Specificity	1–4	> 2	Clarity of idea
Technical validity	1–4	> 1	Feasibility
Innovativeness	1–5	–	Novelty
Competitive advantage	1–4	–	Differentiation
Need validity	0–3	–	User need
Market size	0–3	–	Adoption scale

Table 1: Business-oriented scoring dimensions. The *Threshold* column gives the score cutoff that must be exceeded for downstream dimensions to be scored under the staged screening protocol. "–" indicates a dimension that is only gated by upstream thresholds. Full rubric is provided in Appendix A.

3.2 Product Ideas and Patents

PBIG-DATA contains approximately 300 product ideas generated by LLM-based systems. Each idea is grounded in a patent document and consists of four fields: product title, product description, implementation, and differentiation.

The ideas were produced by multiple independent ideation systems with different prompting and agentic designs, including divergent and convergent ideation, multi-agent discussion, and iterative rewriting pipelines (Yoshiyasu, 2025; Kanumolu et al., 2025; Xu et al., 2025; Terao and Tachioka, 2025; Hoshino et al., 2025; Shimanuki et al., 2025); see Appendix B for an overview. Using outputs from multiple systems reduces dependence on any single generation strategy and helps ensure that disagreement patterns are not artifacts of a specific approach.

The input patents are sampled from the USPTO corpus and span three technical areas: natural language processing (NLP), computer science (CS), and material chemistry (MatChem). Domain diversity is important because feasibility and market judgments may depend on background knowledge and domain-specific assumptions.

3.3 Expert Annotation Protocol

We employ domain experts with both technical and business backgrounds. Across all three domains, we require at least five years of professional experience in the respective domain. Technical experts are additionally required to have a record of publishing peer-reviewed research papers in the domain, while business experts are required to have experience in either consulting or B2B new-business development within the domain. Experts are assigned to domains where they have relevant experience and evaluate only patents they can confidently interpret.

Domain	Experts	Patents	Ideas	Annotations
NLP	12	46	100	1,055
CS	11	48	97	984
MatChem	4	22	110	1,070
Total	–	116	307	3,109

Table 2: Dataset coverage by domain. Annotations count idea–dimension score entries after staged screening.

The annotation process follows a staged screening protocol to avoid forcing scores when a dimension cannot be meaningfully assessed. All experts first score specificity. If an idea does not meet a minimum specificity threshold (see the *Threshold* column of Table 1), downstream dimensions are not scored. On the technical side, technical validity is scored only when specificity passes the threshold, and innovativeness and competitive advantage are scored only when both specificity and technical validity pass their thresholds. On the business side, need validity and market size are scored only when specificity passes the threshold. This protocol produces missing scores by design. We treat missingness as part of the evaluation process rather than annotation noise. The staged screening protocol is illustrated in Appendix C.

Table 2 summarizes the subset of patents and ideas selected for expert scoring in each domain, and the resulting number of score entries after staged screening.

3.4 Problem Formulation for Automatic Judges

Given a patent and a product idea, an automatic judge predicts a score for a target dimension. We compare three judge configurations that correspond to different assumptions about the target signal under heterogeneous expert standards:

- Zero-shot judge: the model predicts scores using only the rubric and task instructions, without access to prior human scoring examples.
- Aggregate judge: the model is conditioned on scoring histories sampled from multiple evaluators, treating pooled behavior as the target.
- Personalized judge: the model is conditioned on scoring history from the target evaluator, aiming to reproduce evaluator-specific standards rather than a pooled consensus.

These configurations allow us to test whether aggregation is an appropriate modeling target when expert disagreement is systematic.

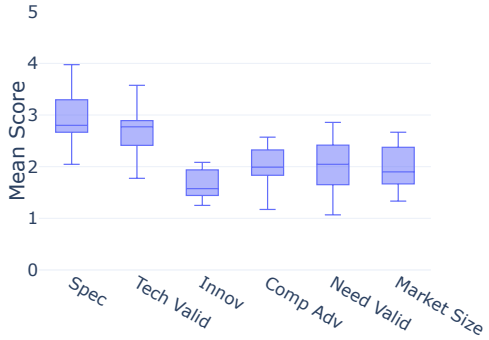

Figure 1: Distribution of per-evaluator mean scores for each dimension.

4 Expert Disagreement in Business Idea Scoring

This section characterizes expert disagreement in PBIG-DATA. The goal is to understand what the expert labels represent as a target for automatic judging. If scores do not converge to a shared ordinal scale, then treating pooled labels as a single ground truth can be a misspecified objective.

4.1 Variation in Evaluator Scoring Scales

We first examine differences in scoring strictness across evaluators. For each evaluator and each dimension, we compute the mean score over the items that the evaluator rated. Figure 1 summarizes these per-evaluator means.

The distributions show substantial inter-evaluator variation for many dimensions. Variation is especially pronounced for need validity and market size, where judgments depend on implicit assumptions about target users and adoption context. This pattern suggests that the same score can reflect different thresholds across evaluators, making direct aggregation of ordinal labels problematic.

4.2 Fine-Grained vs. Coarse Agreement

We quantify agreement at two levels that correspond to different practical uses of scores.

Fine-grained agreement measures whether evaluators agree on ordinal scores. We compute Krippendorff's α (Krippendorff, 2011) for each dimension and domain. Coarse agreement measures whether evaluators select similar subsets of strong ideas even if they disagree on exact scores. For each evaluator and dimension, we define the above-median items

	Fine (α)			Coarse (Jac.)		
Dim.	NLP	CS	Mat.	NLP	CS	Mat.
Spec.	0.06	-0.11	0.04	0.45	0.48	0.45
Tech. val.	-0.03	-0.40	-0.28	0.50	0.33	0.42
Innov.	0.33	0.47	0.46	0.71	0.43	0.54
Comp. adv.	-0.08	0.24	-0.02	0.71	0.33	0.46
Need	-0.23	0.02	0.05	–	–	0.89
Market	0.48	-0.31	0.08	–	–	0.57

Table 3: Expert disagreement summary. Fine-grained agreement is Krippendorff's α. Coarse agreement is mean Jaccard similarity of above-median sets.

(median computed within that evaluator) and compute the mean Jaccard similarity between evaluator pairs that share at least 10 overlapping scored items.

Table 3 reports both metrics. Fine-grained agreement is often close to zero, and in some cases negative, indicating that evaluators do not rank ideas on a shared scale. Coarse agreement is higher across many settings, suggesting that evaluators share some common structure in identifying stronger ideas even when they disagree on exact scores.

These results support two conclusions that matter for judge design. First, disagreement is systematic rather than purely noise, and pooled ordinal labels should not be assumed to represent a stable ground truth. Second, because evaluators can be consistent in coarse selection while differing in score calibration, judge alignment should be evaluated with attention to which target signal is being modeled.

4.3 Implications for Judge Modeling

The disagreement patterns in Table 3 motivate the aggregate versus personalized comparison in the next section. If the correct target is a shared scoring standard, conditioning on mixed evaluator histories should be sufficient. If evaluators apply heterogeneous but internally consistent standards, evaluator-specific conditioning may better reproduce expert behavior. We next test these alternatives using zero-shot, aggregate, and personalized judges.

5 Aggregate vs. Personalized Judges

Section 4 showed that expert scores do not collapse into a shared ordinal scale. This section evaluates how different judge configurations behave under such heterogeneous standards.

5.1 Experimental Setup

We evaluate the zero-shot, aggregate, and personalized judges defined in Section 3.4. The goal is to measure how closely each configuration aligns with expert annotations. We use four Qwen3 models of varying sizes (Qwen Team, 2025): 4B-Instruct-2507, 30B-A3B-Instruct-2507, 30B-A3B-Thinking-2507, and 235B-A22B-Instruct-2507. The 30B-A3B-Thinking model includes reasoning steps. Results for GPT-5 mini are reported in Appendix G. Section 5.3 complements the Krippendorff's α results with coarse selection metrics.

Few-shot examples are drawn from the pool of scored instances in the same domain and same dimension, but always from ideas grounded in different patents than the target to prevent leakage. The personalized judge and the aggregate judge use the same selection logic except for one constraint: for the personalized judge, examples must be scored by the same evaluator as the target; for the aggregate judge, examples must be scored by a different evaluator, creating a cross-evaluator conditioning set. The target item itself is always excluded from the conditioning set. For each shot count, examples are sampled without replacement with three random-seed variants. We note that the evaluation follows a leave-one-out protocol rather than a traditional train/test split. Each target instance is scored by a judge conditioned on examples drawn from the remaining instances in the same domain and dimension but from different patents, so no patent appears simultaneously in the target and in its own conditioning set.

The judge predicts a Likert score together with a self-reported confidence value in the range 0–100 (see Appendix D for the prompt). Following Dong et al. (2024), we discard predictions with confidence below 80. Appendix E shows the discard rates.

Three runs with different random seeds are performed, and majority voting determines the final prediction. Agreement between predictions and expert annotations is measured using Krippendorff's α, consistent with the fine-grained metric in Section 4.2.

5.2 Alignment with Expert Annotations

Figure 2 shows alignment across models and dimensions. For most dimensions and model sizes, personalized judges agree more closely with the corresponding evaluator than aggregate judges. The gap becomes more pronounced for larger models.

Aggregate judges typically outperform zero-shot judges, indicating that historical examples improve calibration. However, aggregate conditioning consistently underperforms personalized conditioning. This suggests that pooled scoring histories encode

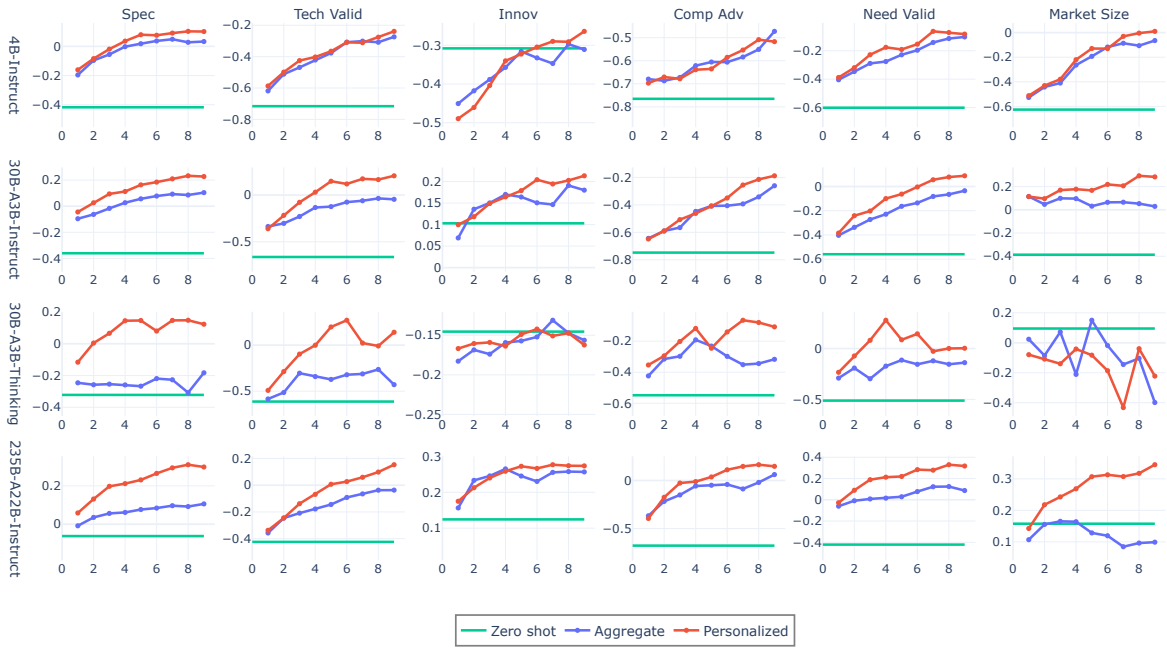

Figure 2: Alignment between automatic judges and expert annotations, measured by Krippendorff's α. X-axis is the number of few-shot examples.

an averaged behavior that does not fully represent any individual evaluator's standard.

Zero-shot performance remains near zero in many dimensions, demonstrating that the rubric alone does not define a unique scoring scale. Historical examples are necessary for calibration, but the type of history matters. Appendix F presents a qualitative case study in which aggregate and personalized judges assign different innovativeness scores to the same idea, illustrating how evaluator-specific calibration manifests in individual predictions.

5.3 Coarse Selection Metrics

Section 4.2 argued that expert disagreement is more pronounced at the fine-grained ordinal level than at the coarse selection level. To check whether the personalization advantage also extends to coarse selection behavior, we report two additional metrics computed on the 9-shot judges (the maximum shot count in our experiments).

Above-median Jaccard similarity. For each evaluator and dimension, we compute the Jaccard similarity between the evaluator's and the judge's above-median item sets (Table 4). Personalized judges achieve higher Jaccard similarity than aggregate judges on five of the six dimensions.

Top-50% overlap. We also compute the fraction of each expert's top-50% items that the judge also ranks in its own top-50% (Table 5). Personalized judges again dominate on most dimensions, with particularly large gains on technical validity and market size.

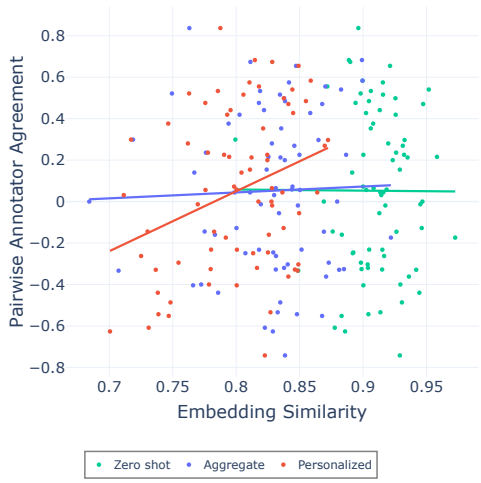

Figure 3: Relationship between evaluator agreement and similarity of judge-generated reasoning texts. A positive trend appears only under personalized conditioning.

Together, these metrics confirm that the personalization advantage extends beyond fine-grained ordinal agreement to coarse selection behavior, especially on dimensions where expert disagreement is highest.

Dim.	Aggregate	Personalized
Spec.	0.747	0.780
Tech. val.	0.594	0.714
Innov.	0.759	0.715
Comp. adv.	0.718	0.725
Need	0.483	0.574
Market	0.567	0.648

Table 4: Above-median Jaccard similarity between judge and expert selections, at 9-shot.

Dim.	Aggregate	Personalized
Spec.	0.594	0.638
Tech. val.	0.455	0.621
Innov.	0.681	0.560
Comp. adv.	0.660	0.638
Need	0.611	0.619
Market	0.486	0.667

Table 5: Top-50% overlap: fraction of each expert's top-50% items that the judge also ranks in its top 50%, at 9-shot.

5.4 Evaluator Agreement and Reasoning Structure

Beyond score alignment, we examine whether personalized judges also capture differences in evaluators' underlying reasoning patterns (Figure 3). For evaluator pairs with overlapping idea–dimension annotations, we compute two quantities. First, we measure their empirical agreement using mean Krippendorff's α across shared instances. Second, we compute the cosine similarity between reasoning texts generated by the zero-shot, aggregate, and personalized judges. Embeddings are generated by Qwen3-Embedding-8B (Qwen Team, 2025). Under personalized conditioning, a positive relationship emerges between evaluator agreement and reasoning similarity (Pearson $r = 0.31$). Evaluators who agree more on scores tend to produce more semantically similar reasoning when modeled individually. In contrast, this relationship is near zero under aggregate or zero-shot configurations.

This result indicates that personalized judges do not merely replicate score distributions. They capture evaluator-specific evaluation policies that manifest both in numeric scores and in textual justifications. When two evaluators share similar standards, their personalized judges generate correspondingly similar reasoning. When standards diverge, reasoning diverges as well. These findings strengthen the interpretation that disagreement in PBIG-DATA reflects structured heterogeneity rather than noise. Aggregate conditioning smooths differences, while personalized conditioning preserves them.

6 Conclusion

This paper examined business idea evaluation under heterogeneous expert standards. Using PBIG-DATA, we showed that expert scores often fail to converge to a shared ordinal scale, even when a common rubric is provided. Yet, agreement under coarse selection indicates that disagreement at the fine-grained level reflects calibrated but distinct evaluation policies rather than random noise.

We compared zero-shot, aggregate, and personalized judge configurations under these conditions. Aggregate judges improve over rubric-only prompting but approximate an averaged consensus that does not fully represent any individual evaluator. Personalized judges consistently achieve higher alignment with the corresponding evaluator and preserve evaluator-specific reasoning structure. These findings suggest that, in pluralistic evaluation settings, the choice of target signal is not neutral: modeling a pooled standard and modeling individual standards lead to substantively different behaviors.

From an industry perspective, these results have practical implications for deploying LLM-based evaluation in ideation workflows. In real-world organizations, business ideas are often assessed by stakeholders with different roles and incentives. Collapsing their judgments into a single aggregate score may obscure meaningful differences in perspective. Rather than enforcing artificial consensus, evaluation systems can surface structured disagreement and assist decision-makers in navigating it.

More broadly, our results indicate that evaluation for business ideation cannot be reduced to a single ground truth without losing information about stakeholder-specific judgment. Future work should explore evaluation protocols that explicitly account for heterogeneous standards, including calibration mechanisms and multi-perspective scoring that preserve rather than suppress evaluator diversity.

Ethical Considerations and Licensing

All data in PBIG-DATA are derived from public patent documents and automatically generated ideas submitted to an open shared task. No personal or sensitive information is included. Annotations for the computer science and NLP domains were performed by consenting professional annotators and are released at https://stockmarkteam.github.io/pbig-data/. For

the materials chemistry domain, annotations were conducted in collaboration with a private company, and their release will follow contractual.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354. Association for Computational Linguistics.

Curtis R Carlson and William W. Wilmot. 2006. *Innovation: The Five Disciplines for Creating What Customers Want*. Crown Business.

Robert G. Cooper. 1990. Stage-gate systems: A new tool for managing new products. *Business Horizons*, 33(3):44–54.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a Personalized Judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95. Association for Computational Linguistics.

Wataru Hirota, Chung-Chi Chen, Tomoko Ohkuma, Tomoki Taniguchi, and Tatsuya Ishigaki. 2025. Overview of PBIG shared task at AgentScen 2025: Product business idea generation from patents. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 35–42.

Mizuki Hoshino, Shun Shiramatsu, and Fuminori Nagasawa. 2025. A business idea generation framework based on creative multi-agent discussions. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 82–85.

Gopichand Kanumolu, Ashok Urlana, Vinayak Kumar Charaka, and Bala Mallikarjunarao Garlapati. 2025. Agent ideate: A framework for product idea generation from patents using agentic AI. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Masaya Shimanuki, Naoto Shimizu, Kentaro Kinugasa, and Hiroki Sugisawa. 2025. Team mcg dsn at the agentscen shared task: Knowledge integration and self-improvement via LLMs for generating business ideas from patent documents. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 86–92.

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

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. JudgeBench: A Benchmark for Evaluating LLM-based Judges. In *Proceedings of The Thirteenth International Conference on Learning Representations*.

Yasunori Terao and Yuuki Tachioka. 2025. Collaborative invention: Team ditlab at the AgentScen shared task – refining patent-based product ideation via LLM-guided selection and rewriting. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 67–81.

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

Keisuke Ueda, Wataru Hirota, Kosuke Takahashi, Takahiro Omi, Kosuke Arima, and Tatsuya Ishigaki. 2025. Exploring the design of multi-agent LLM dialogues for research ideation. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 322–337. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2410.21819*.

Yuzheng Xu, Tosho Hirasawa, Seiya Kawano, Shota Kato, and Tadashi Kozuno. 2025. MK2 at PBIG competition: A prompt generation solution. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 58–66.

Hayato Yoshiyasu. 2025. Team NS_NLP at the AgentScen shared task: Structured ideation using divergent and convergent thinking. In *Proceedings of the 2nd Workshop on Agent AI for Scenario Planning*, pages 43–49.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852. Association for Computing Machinery.

A Full Scoring Rubric

This section provides the complete rubric used for expert scoring. The rubric was designed to reflect practical business-oriented evaluation rather than purely linguistic quality. Each dimension corresponds to a distinct aspect of product assessment that may be emphasized differently by different evaluators.

The six dimensions separate technical feasibility (specificity and technical validity), novelty and differentiation (innovativeness and competitive advantage), and commercial relevance (need validity and market size). Need validity and market size are scored on a 0–3 scale in our rubric.

The rubric defines discrete score levels with written descriptions to encourage consistent interpretation. However, as shown in the main text, even under a shared rubric, experts apply different implicit assumptions and thresholds. We therefore release the full rubric to enable reproducibility and to allow future work to investigate alternative judge designs under the same evaluation specification.

B Overview of Idea Generation Systems

This section briefly summarizes the ideation systems whose outputs are included in PBIG-DATA. The purpose of this overview is to clarify data provenance and diversity, not to compare generation performance. The dataset includes ideas generated by multiple independent systems (Yoshiyasu, 2025; Kanumolu et al., 2025; Xu et al., 2025; Terao and Tachioka, 2025; Hoshino et al., 2025; Shimanuki et al., 2025). These systems vary in prompting strategies, use of multi-agent interaction, and refinement mechanisms. Some rely primarily on structured prompting, while others employ multi-agent discussions, critique–revision loops, or iterative rewriting pipelines.

Generation performance comparisons and system-level analyses are reported in prior workshop publications (Hirota et al., 2025). In contrast, the present work does not evaluate or rank these systems. Their outputs are used collectively to analyze evaluation behavior under heterogeneous expert standards. Using ideas from multiple systems ensures that the observed disagreement patterns are not artifacts of a single generation strategy. The conclusions of this paper therefore concern evaluation methodology rather than generation quality.

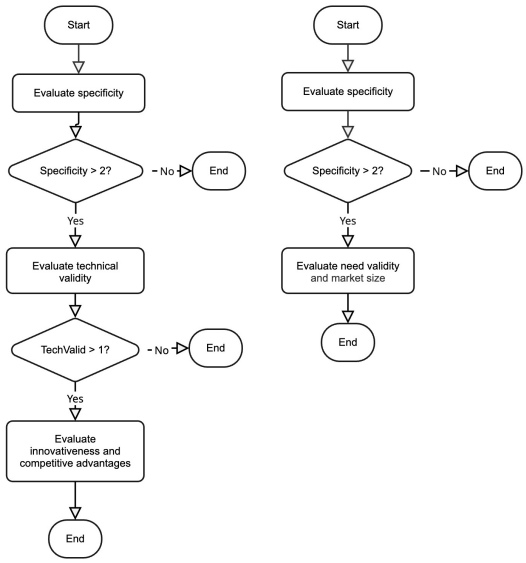

Figure 4: Staged screening protocol for expert scoring.

C Staged Screening Protocol

Figure 4 shows the staged screening protocol used during expert scoring. Downstream dimensions are evaluated only when prerequisite conditions are met, which produces missing scores by design.

D Judge Prompt Schema

All judge configurations use the same prompt structure, shown in Figure 5.

E Confidence Filtering Discard Rates

Predictions whose self-reported confidence value is below 80 are discarded, as described in Section 5. Table 7 reports the proportion of discarded predictions per dimension and domain. All discard rates are below 0.32%, so the filtering step has a negligible effect on the reported alignment patterns.

F Qualitative Case Study

To illustrate how aggregate and personalized judges can diverge even for the same item, we examine an innovativeness assessment of an idea that applies a patented UI-customization technology to electronic medical record (EMR) systems with role-based permissions. The target evaluator assigned a conservative innovativeness score of 2, reflecting the view that role-based EMR customization, while a plausible extension, is a known use case within healthcare IT. The personalized judge reproduced this score and justified it by noting that "*the idea*

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	Feasibility of implementing the idea using the 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	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.	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.
Market size	Number of potential users.	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 companies.

Table 6: Full rubric for the six business-oriented scoring dimensions.

Dim.	NLP	CS	Mat.
Spec.	0.012%	0.103%	0.319%
Tech. val.	0.026%	0.008%	0.108%
Innov.	0.000%	0.000%	0.000%
Comp. adv.	0.001%	0.003%	0.006%
Need	0.004%	0.000%	0.003%
Market	0.000%	0.000%	0.000%

Table 7: Proportion of predictions discarded by confidence filtering (confidence < 80), by dimension and domain.

Dim.	Aggregate	Personalized
Spec.	-0.083	0.084
Tech. val.	-0.042	-0.012
Innov.	0.235	0.233
Comp. adv.	-0.028	-0.119
Need	0.189	0.404
Market	0.048	0.257

Table 8: Krippendorff's α between gpt-5-mini-2025-08-07 judge predictions and expert annotations, under aggregate and personalized conditioning.

applies the patented UI customization technology to EMR systems with role-based permissions, which is a known use case in healthcare IT, but the specific implementation for dynamic clinical workflows is not yet widely explored." The aggregate judge, conditioned on mixed-evaluator histories, instead returned a score of 4 and framed the same idea more generously as "*a plausible but not obvious extension of the patent … beyond generic customization by integrating into clinical workflows and compliance needs*." The two judges thus agree on the qualitative description of the idea but apply different innovation thresholds. Under personalized conditioning, the judge inherits the target evaluator's conservative calibration; under aggregate conditioning, the mixed history pulls the score toward a more permissive pooled standard that no individual evaluator necessarily holds.

G Results for GPT-5 mini

To test whether the personalization advantage generalizes beyond the Qwen3 family, we repeat the main judge comparison with gpt-5-mini-2025-08-07. Table 8 reports Krippendorff's α between the judge predictions and expert annotations for the aggregate and personalized configurations. Personalized conditioning yields alignment that is higher than or comparable to the aggregate judge on five of the six dimensions, with the largest gaps on need validity and market size. This mirrors the pattern observed for Qwen3 models in Section 5 and supports the claim that the personalization advantage is not an artifact of a single model family.

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 corresponding score>
<Example 2: patent, idea, and corresponding score>
...
<Example N: patent, idea, and corresponding 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>}
```

Figure 5: The prompt template for LLM-as-a-Judge models. Angle brackets (<...>) denote placeholders.