Kaleidoscope

AI-Powered Innovation Intelligence Engine

Carnegie Mellon University, Tepper School of Business
Master of Science in Business Analytics, 2026

Eddie Li
Chaitanya Kumar Soma
Gitanjali Roy
Yijia Pan

Corporate Sponsor: Honda Research Institute (HRI) & 99P Labs
Project Advisor: Professor Deepak Agrawal
April 2026

TABLE OF CONTENTS

PROJECT OVERVIEW

Organizations that rely on large language models for strategic brainstorming face a problem that is subtle but consequential: the AI produces outputs that look diverse but are not. Portfolios generated through unconstrained prompting contain the same underlying concepts dressed in different words, making it impossible for reviewers to determine which ideas genuinely represent distinct strategic directions and which are paraphrases of ideas already in the list. For Honda Research Institute, a company actively exploring the future of autonomous mobility across robotaxi, eVTOL, and connected vehicle platforms, this is a direct cost in analyst time, strategic clarity, and the quality of decisions that flow downstream from the ideation process.

Kaleidoscope is our response to this problem. It is a multi-agent AI system that treats innovation portfolio construction as a selection and optimization problem over semantic space rather than a generation problem. The system decomposes every candidate idea into six comparable dimensions, embeds each dimension independently, and computes weighted cosine similarity between incoming candidates and the existing portfolio before deciding whether to accept or reject each idea. The result is a portfolio in which every accepted idea is provably distinct from every other accepted idea along the dimensions that matter most for strategic differentiation.

The three-agent architecture automates the full workflow from topic input to deliverable. The Explorer Agent generates and screens candidates. The Refiner Agent expands the highest-novelty ideas into full concept briefs ready for decision-making. The Patent Validation Agent cross-checks top ideas against live patent data to surface early intellectual property risk signals. All outputs flow into a structured four-tab Excel portfolio that documents the idea and the evidence behind their selection. This makes the ideation process auditable and repeatable in a way that unstructured brainstorming won't be able to reach.

This report documents the full methodology, the experimental program that validated it, the results, and the takeaway through building and iterating on the system across five versions of the core generation architecture.

BACKGROUND & BUSINESS PROBLEM

The Problem Honda Brought to This Project

Honda Research Institute's challenge was LLM outputs look diverse but are not. The same concept, repeated in different words, appeared throughout portfolios generated through standard prompting approaches, with no way to detect or correct the redundancy without semantic measurement tools that did not exist in the existing workflow.

To understand why this happens, it helps to understand how large language models generate text. These models assign probability distributions over possible next tokens based on patterns learned from training data. When asked to brainstorm ideas about autonomous vehicles, the model assigns the highest probability to the most statistically common outputs in its training distribution. This means that it gravitates toward the concepts, framings, and solution types that appear most frequently in the literature it learned from. The result is not random but highly clustered around the center of the distribution. Ideas that appear diverse at the vocabulary level are often semantically identical because they describe the same underlying concept through different surface expressions.

This clustering behavior manifests in three specific failure modes that we identified and measured in the unstructured baseline portfolio. First, invisible redundancy. Ideas with completely different vocabulary are semantically identical when measured in embedding space. In our 200-idea baseline test on Honda robotaxi strategy, we found that a substantial number of idea pairs exceeded the 0.70 cosine similarity threshold that defines near-duplication. This is invisible to any human reviewer who evaluates ideas based on their surface wording. Second, category collapse. Unstructured generation defaults to the most statistically probable value categories. In the same 200-idea baseline, one value type dominated the portfolio at 16.5%, crowding out categories that might represent genuinely differentiated strategic directions. Third, radical ideas absent. Only about 3% of ideas generated through standard prompting are genuinely transformative or cross-industry in nature. The model's probability bias toward familiar patterns leaves the tail of the distribution, where the most surprising and potentially valuable ideas live.

The fourth failure mode, which underlies all the others, is the absence of a measurement system. Without embedding-based similarity scoring, none of these patterns are visible. There is no way to know the portfolio was redundant, no way to quantify the degree of redundancy, and no way to know whether any corrective action was working. Measurement is the prerequisite for improvement, and this is what Kaleidoscope addresses first.

Why This Matters for Honda

Honda Research Institute operates at the intersection of multiple high-stakes strategic questions simultaneously. The company is evaluating robotaxi deployment strategies, eVTOL market entry, autonomous mobility infrastructure, and the role of AI in vehicle systems. These are all domains where the competitive landscape is evolving rapidly and the cost of missing a genuinely novel strategic direction is high. In this context, an ideation process that systematically gravitates toward the center of the distribution is strategically dangerous.

The business cost of redundancy in an R&D ideation context takes several forms. First, expert review time is spent on ideas that do not represent distinct strategic choices. When an analyst or leadership team evaluates a portfolio that contains many near-duplicate pairs, they spend cognitive effort distinguishing concepts that differ only in phrasing, not in substance. Second, genuinely novel ideas are underrepresented relative to their strategic value. The transformative ideas that emerge from unstructured generation are buried under the volume of incremental variations, making it harder to surface and prioritize them. Third, the ideation process cannot be repeated reliably. Because unstructured prompting produces different

redundancy profiles on each run with no way to audit or compare them, there is no foundation for building institutional knowledge about which conceptual territories have been explored and which remain open.

Kaleidoscope addresses all three of these costs directly. By reducing duplicate pairs to near-zero, it ensures that every idea in the portfolio represents a distinct strategic choice worth evaluating. By enforcing structural diversity rules that increase the share of radical and cross-industry ideas, it shifts the distribution toward the tail where the most valuable ideas are most likely to appear. And by producing auditable outputs with documented novelty scores and dimension labels, it creates a foundation for repeatable and comparable ideation runs across topics and time periods.

PROJECT SCOPE AND LIMITATIONS

This project focuses on building and validating a structured AI ideation system for Honda Research Institute that addresses the specific failure modes identified in unstructured LLM-generated portfolios. The scope encompasses the design and implementation of the multi-agent pipeline, the experimental program comparing structured and unstructured generation across multiple configurations, and the integration of patent validation as an early-stage IP risk screening layer.

The system is designed to operate as a decision support tool, not as a replacement for human strategic judgment. Every number it produces (including novelty scores, duplicate counts, patent similarity scores) is a signal for human review. Patent outputs are explicitly framed as triage signals for early screening and workshop prioritization. They do not constitute claims analysis or freedom-to-operate assessment, both of which require legal expertise and access to comprehensive patent databases that extend beyond what is available through the SerpAPI and Google Patents retrieval path used in this implementation.

The patent retrieval architecture uses Google Patents data accessed via SerpAPI. This provides a broad and reasonably current view of the patent landscape but does not offer exhaustive USPTO or global coverage. The system is designed to be modular, so additional retrieval adapters for USPTO APIs, EPO, and WIPO can be integrated behind the same scoring interface as those become available and compliant with applicable licensing terms. For the purposes of this project, the Google Patents path is sufficient for early-stage risk screening while the architectural foundation is established.

The dimension weights used in the weighted similarity calculation were derived empirically from the 200-idea baseline library generated on Honda robotaxi strategy topics. These weights reflect the discriminating power of each dimension in that specific domain. When the framework is applied to substantially different domains (like healthcare, energy infrastructure, consumer software) the weights should be recalibrated through a similar empirical analysis of a domain-specific idea library. The current weights are a reasonable starting point for mobility-adjacent domains but should not be treated as universal.

The experimental program covers five versions of the structured generation architecture tested against an unstructured baseline, with all comparisons made on a common weighted novelty scale using the same embedding model and formula. The results are specific to the Honda robotaxi and eVTOL strategy topics tested and are presented as empirical evidence for the system's effectiveness rather than as universal claims about LLM behavior in all domains. Generalizability across domains is an area for future validation.

DATA SOURCE

The primary data source for the experimental program is a corpus of ideas generated through both structured and unstructured prompting on Honda robotaxi and eVTOL strategy topics, using GPT-4o-mini at temperature 1.0 as the generation model. The unstructured baseline consists of 200 ideas generated through a single bulk prompt with no diversity constraints, which forms the performance floor against which all structured configurations are measured.

Each idea in both the baseline and structured portfolios is represented as a structured record containing six labeled dimensions including core mechanism, problem opportunity, intended outcome, stakeholder, context, and value type, plus the narrative idea text itself. The dimension labels are generated through a separate GPT call that extracts each field concisely and consistently. The labeling prompt instructs the model to keep each field to eight words or fewer and to reuse the same label when ideas address the same concept, ensuring that the embedding-based similarity calculation reflects genuine conceptual overlap rather than surface variation in how dimensions are described.

Embeddings are generated using OpenAI's text-embedding-3-small model at 1536 dimensions. Each idea is embedded both at the whole-idea level for novelty analytics and duplicate detection, and at the individual dimension level for the weighted similarity gating that drives the selection loop during structured generation. This dual embedding approach allows the system to apply different weights to different aspects of an idea when computing similarity, rather than treating the idea as an undifferentiated block of text.

For the patent validation component, data is retrieved in real time from Google Patents via SerpAPI at the time each run is executed. The query is constructed from the idea's core mechanism and the broader topic string, and the top patents returned are embedded and scored against the idea embedding. No patent data is stored between runs. Each patent check is a fresh retrieval against the live patent database.

PROPOSED SOLUTION & METHODOLOGY

Kaleidoscope's methodology rests on a single conceptual shift: separating the generation of ideas from the selection of ideas. In standard LLM-based brainstorming, generation and selection are conflated. The model generates, and whatever comes out is the portfolio. Kaleidoscope makes selection an measurable and algorithmic step that operates independently of generation. The AI generates freely within the topic space and the framework decides what enters the portfolio based on semantic diversity criteria applied through embedding-based similarity measurement.

This separation is what makes the framework tool-agnostic. The decomposition into six dimensions, the weighted hierarchy, and the selection logic can be applied manually on a whiteboard with domain experts evaluating ideas against previously accepted ideas. The AI pipeline is one implementation of this framework. The insight that not all differences between ideas are equally meaningful, and that a weighted hierarchy of dimensions can capture this intuition formally, is independent of any technology stack.

Step 1: Idea Representation and Dimension Labeling

Every candidate idea in the Kaleidoscope framework is decomposed into six labeled dimensions before any similarity calculation occurs. These dimensions are: core mechanism, which captures the primary technical or strategic mechanism the idea employs; problem opportunity, which captures the specific gap, pain point, or unmet need the idea addresses; intended outcome, which captures the desired result the idea is trying to achieve; stakeholder, which identifies the primary user, customer, or organizational actor the idea affects; context, which describes the setting, touchpoint, or deployment environment where the idea applies; and value type, which classifies the type of value created, such as convenience, safety, efficiency, or trust.

This decomposition serves two purposes. First, it enables weighted similarity calculation across dimensions that have different discriminating power. Two ideas can share a core mechanism but differ entirely in their stakeholder and intended outcome; the weighted formula can capture this as genuinely distinct while two ideas that share mechanism, stakeholder, and outcome are appropriately flagged as near-duplicates even if their surface wording is completely different. Second, it makes the similarity calculation interpretable. When a candidate idea is rejected, the output logs which dimension drove the rejection and what the similarity score was, giving the team a clear picture of why the portfolio looks the way it does.

The dimension weights are derived empirically from the baseline 200-idea library by counting the number of unique values per dimension. A dimension with more unique values across the library is more useful for telling ideas apart and therefore carries more weight in the similarity formula.

Dimension	Weight	Rationale
Core Mechanism	30%	Every idea has a unique mechanism. Strongest discriminating signal
Problem / Opportunity	25%	Almost every idea targets a different problem space
Intended Outcome	20%	Two ideas can share a mechanism but aim for different results
Stakeholder	12%	Meaningful variation but not the primary differentiator
Context	8%	High count but mostly surface-level setting variation
Value Type	5%	Only 15 categories. Intentionally low weight to avoid over-penalization

Core mechanism has the highest weight at 30% because every idea in the baseline had a unique mechanism, making it the strongest discriminating signal. Value type has the lowest weight at 5% because only 15 categories exist across the entire library, making it the weakest discriminating signal.

Step 2: Embedding and Weighted Similarity

All embeddings use OpenAI's text-embedding-3-small model at 1536 dimensions. The choice of this model over whole-idea embedding approaches reflects a deliberate design decision: embedding each dimension separately allows the similarity calculation to be sensitive to what kind of difference exists between two ideas, not just whether a difference exists. Two ideas that describe the same mechanism in different syntactic forms will have high cosine similarity on the mechanism dimension regardless of how different their other dimensions are; the weighted formula will then correctly flag them as redundant on the dimension that matters most.

The weighted cosine similarity formula operates as follows. For each dimension, the embedding of that dimension's text from idea A and idea B is computed, and the cosine similarity between the two embeddings is calculated. These per-dimension similarities are then weighted by the empirically derived weights and summed to produce the overall weighted similarity score. This score ranges from zero to one, where zero indicates completely orthogonal ideas in weighted embedding space and one indicates perfect semantic identity across all dimensions with their respective weights applied.

$$\textit{Weighted Similarity} = 0.30 \times \textit{sim(mechanism)} + 0.25 \times \textit{sim(problem)} + 0.20 \times \textit{sim(outcome)} + 0.12 \times \textit{sim(stakeholder)} + 0.08 \times \textit{sim(context)} + 0.05 \times \textit{sim(value type)}$$

Novelty for each accepted idea is then defined as 1 minus the average weighted similarity to its five nearest neighbors in the accepted library.

The duplicate detection threshold used in the Diversity Snapshot is 0.70 on whole-idea embeddings. This is a deliberately conservative threshold that flags only strong near-duplicates. The generation gate threshold is separate and operates on weighted dimension similarity rather than whole-idea similarity; the default is 0.65 and the optimized configuration uses 0.55. These thresholds are governance parameters that should be locked for production studies and documented in run outputs so that differences between runs can be explained by configuration differences rather than attributed to randomness.

Step 3: Structured Generation and Selection Loop

The Explorer Agent generates candidate ideas one at a time. Each candidate is immediately embedded across all six dimensions and scored against every idea already in the accepted library using the weighted similarity formula. If the candidate's maximum weighted similarity to any accepted idea falls below the gate threshold, it is accepted into the portfolio. If it exceeds the threshold, it is rejected and a new candidate is generated in its place. This continues until the portfolio reaches the target size, with each acceptance decision made in real time against the current state of the library.

Step 4: Refiner Agent — Concept Brief Expansion

The Refiner Agent takes the highest-novelty ideas from the accepted library and expands each into a full concept brief structured for strategic decision-making. The brief format covers six components: problem statement, which describes the pain point, unmet need, or business opportunity in one to two sentences without describing how the idea solves it; proposed solution, which explains the mechanism and value creation logic in concrete terms; target user, which identifies the primary stakeholder and their specific situation and constraints; business model, which outlines how Honda could monetize or deploy the concept; key risks, which identifies the most significant technical, regulatory, and market obstacles; and Honda advantage, which explains why Honda is specifically positioned to execute this idea relative to other players.

The Refiner operates in parallel across all briefs, generating them simultaneously to minimize total runtime. Each brief is generated with a lower temperature than the Explorer (0.3 versus 1.0) to prioritize consistency and coherence over creativity (the creative work is done by the Explorer); the Refiner's job is to make the best ideas legible and actionable for a non-technical audience. The briefs are written at a level that a business stakeholder without deep technical background can evaluate, which means avoiding jargon where possible and grounding every claim about business model and Honda advantage in the specific context of the idea rather than generic statements about AI or innovation.

Step 5: Patent Validation Agent

The Patent Validation Agent cross-checks the top ideas from the Refiner output against live patent data retrieved via SerpAPI. The retrieval query is constructed from the idea's core mechanism label and the broader topic string, designed to surface patents that address the same technical territory as the idea rather than patents that merely share vocabulary. Up to five patents are retrieved per idea, and each patent's title and abstract are embedded using the same text-embedding-3-small model used throughout the system.

Cosine similarity is then computed between the idea embedding and each patent embedding. This embedding-based approach is a deliberate improvement over keyword-based or TF-IDF similarity methods that were explored in the initial research notebook. Keyword methods fail when an idea and a patent describe the same technical mechanism using different vocabular, which is exactly the situation most likely to produce a genuine intellectual property risk. Embedding-based similarity catches semantic near-matches regardless of surface vocabulary, which makes it a more reliable early-stage screening tool.

Ideas whose closest patent similarity falls below 0.35 are fast-pathed to a Low-risk classification. For ideas above this threshold, a GPT call reads the idea text and the top one to three matching patent abstracts and produces a structured risk assessment: a risk level of High, Medium, or Low based on the similarity profile and the specificity of the overlap, plus a one-sentence plain-English explanation of the assessment grounded in the actual patent text. The risk thresholds are: below 50% similarity indicates Low risk, between 50% and 70% indicates Medium risk warranting refinement of the idea's core mechanism, and above 70% indicates High risk suggesting the idea may need to pivot its mechanism substantially before being viable for patent filing.

All patent outputs carry the explicit caveat that they are triage signals for early screening, not legal clearance. The Google Patents via SerpAPI retrieval path does not provide exhaustive coverage of the global patent landscape, and similarity scores should not be interpreted as clearance from freedom-to-operate risk. Claims-level analysis, prior art searching, and freedom-to-operate assessment require legal expertise and access to comprehensive databases that are outside the scope of this system.

EXPERIMENTAL DESIGN

The experimental program is structured around a comparison between two generation architectures applied to the same topic, the same model, and the same temperature. Group A is the unstructured baseline: 200 ideas generated through a single bulk prompt with temperature 1.0, no constraints, no semantic screening, and no rejection mechanism. The AI sees no history. It generates the entire portfolio in one shot. Group A establishes the performance floor and documents the failure modes that the structured approach is designed to correct.

Group B is the structured approach, implemented across five versions that each corrected the failure mode found in the previous version. All Group B versions use temperature 1.0 during generation to maintain comparability with Group A on the generation side; the diversity enforcement is applied through the selection mechanism and the 14-rule prompt framework rather than by changing the generation temperature. This design ensures that differences in outcome can be attributed to the structural differences in the architecture rather than to differences in generation randomness.

Group A — Unstructured	Group B — Structured
200 ideas, single bulk prompt	200 ideas, one at a time with 14 rules
Temperature 1.0, no constraints	Real-time weighted dimension similarity gate
No semantic screening or rejection	Semantic rejection before acceptance
One-shot — AI sees no history	AI sees accepted history — anchored generation

All versions are measured on a common set of metrics computed using the same embedding model and the same formulas. The primary metric is weighted novelty score, computed as described in the methodology section using the six-dimension weighted formula. Secondary metrics include near-duplicate pair count (ideas whose nearest-neighbor cosine similarity on whole-idea embeddings exceeds 0.70), novelty label distribution (share of ideas classified as High, Medium, and Low novelty based on portfolio percentile thresholds), share of radical or transformative ideas, and share of cross-industry ideas. Statistical significance is assessed using both parametric (t-test) and non-parametric (Mann-Whitney U) tests to ensure the conclusions are robust to distributional assumptions.

The version history for Group B reflects an iterative discovery process. Version 1 encoded all 13 diversity rules in a single prompt and relied on the language model to comply with them during generation. This approach failed because LLM compliance with multi-constraint prompts is unreliable at scale; only 71% of generated ideas actually followed the rules, producing a portfolio of 133 ideas rather than the target 200. The lesson was that rule enforcement cannot be delegated to the generation model; it must be implemented in the selection mechanism.

Version 2 introduced balanced value type caps enforced through the selection mechanism rather than through the prompt, which successfully eliminated duplicate pairs. However, weighted novelty remained below the Group A baseline, suggesting that eliminating redundancy through capping alone is insufficient to improve novelty — it removes the worst ideas but does not actively push the portfolio toward higher-novelty territory. Version 3 introduced the weighted dimension similarity gate, which replaced the simple caps with a comprehensive similarity check against the entire accepted library before each acceptance decision. This version achieved both zero duplicate pairs and an 8.5% novelty improvement over Group A at 200 ideas, establishing the architecture that all subsequent versions build on.

Version 4 tested whether the selection mechanism alone was sufficient without the 14-rule diversity prompt framework, by allowing free generation and checking similarity invisibly after the fact. This version produced the worst weighted novelty of any configuration tested, confirming that generation-side diversity enforcement and selection-side similarity gating are both necessary components rather than substitutes. Version 5 explored batch selection, generating ten candidates and selecting the most novel rather than accepting the first candidate that passed the gate. This approach worked well initially but degraded severely after approximately idea 150, which pointed toward the domain saturation phenomenon that motivated the final configuration.

Version	Key Change	Outcome	Status
v1	13 rules in a single prompt	LLM compliance collapsed — only 71% of ideas followed rules	Failed
v2	Balanced value type caps	Zero duplicate pairs achieved but weighted novelty lower than Group A	Partial
v3	Weighted dimension similarity gate	+8.5% novelty over Group A. Zero duplicate pairs.	Best at 200 ideas
v4	Free generation with invisible check	Worst weighted novelty (0.4156) — removing structure caused collapse	Failed
v5	Batch selection from 10 candidates	Domain saturation at ~idea 150 — underperformed unstructured baseline	Partial

RESULTS & ANALYSIS

Novelty and Redundancy Outcomes

The core quantitative finding of the experimental program is that the novelty-redundancy tension is resolved at 100 ideas with a similarity gate threshold of 0.55. The optimized configuration achieves a weighted novelty score of 0.4912 against the Group A unstructured baseline of 0.4962, a gap of 0.005 or 1.0% below baseline. At the same time, it reduces near-duplicate pairs from 129 in the unstructured baseline to a single pair — a 99.6% reduction.

This result is significant because it demonstrates that the two objectives are not fundamentally in conflict. The apparent trade-off observed in earlier versions of the architecture, where eliminating redundancy came at the cost of reduced novelty, was a consequence of operating the system in the domain saturation zone above 100 ideas, not a fundamental property of the structured approach. Within the pre-saturation zone, the 14-rule framework achieves both high novelty and near-zero redundancy simultaneously.

The shift in novelty label distribution is equally striking. The share of Low Novelty ideas falls from 46% in Group A to 1% in the optimized Group B configuration — a 98% reduction. The share of High Novelty ideas rises from 20% to 30% — a 27% increase. The share of ideas classified as radical or transformative rises from approximately 3% in Group A to 84.5% in the best Group B version — a 28-fold increase. Cross-industry ideas, which appear at roughly 8% frequency in unstructured generation, achieve universal coverage in the structured portfolio. These distribution shifts are consistent with the design intent of the 14-rule framework, which explicitly targets radical and cross-industry ideas as required components of the portfolio.

The statistical tests confirm that these differences are not due to chance. The t-test comparing the Group A and Group B v3 novelty distributions produces a t-statistic of 4.63 with a p-value of 0.000005. The Mann-Whitney U test produces a p-value below 0.000001. Both results are significant at p less than 0.05 with a large effect size, confirming that the structured approach produces a measurably and reliably different distribution of novelty scores, not a random variation around the same underlying distribution.

Domain Saturation Discovery

The most important empirical discovery from the experimental program was the domain saturation phenomenon. Version 5 of the Group B architecture, which used batch selection from ten candidates per round, performed well in the early portion of each run but degraded sharply after approximately idea 150. Weighted novelty scores fell and duplicate pair counts rose in the same region of the run, indicating that the embedding space available for the topic was becoming exhausted regardless of the selection strategy.

This finding motivated two changes to the final configuration. First, the target portfolio size was reduced from 200 to 100 ideas, keeping the system in the pre-saturation zone where the semantic space is still rich enough to support genuinely distinct ideas at each selection step. Second, the gate threshold was tightened from 0.65 to 0.55, which forces the generation loop to search harder for novel candidates before accepting each one and counteracts the natural tendency of later rounds to accept ideas that are marginally novel rather than genuinely distinct.

The saturation discovery has a practical implication beyond the immediate experimental context: the optimal portfolio size for any given topic is a function of the breadth of that topic's embedding space, not a fixed number that can be applied universally. Narrow, highly specific topics will saturate earlier than broad, cross-domain topics. Future work should explore adaptive stopping criteria that monitor novelty trajectories and trigger the end of a run when marginal novelty per accepted idea falls below a defined threshold.

Clustering Analysis

K-means clustering of the 200-idea baseline library across k values from 5 to 15 produces silhouette scores between 0.037 and 0.042. These modest absolute values are consistent with a dense semantic cloud of related mobility ideas — the embedding space does not contain well-separated, internally cohesive clusters of the kind that would be expected if the ideas naturally divided into distinct conceptual categories. Instead, the ideas form a continuous manifold with gradients of similarity rather than discrete islands.

This finding reinforces the case for explicit selection using weighted dimension similarity rather than relying on cluster structure to enforce diversity. If the embedding space were naturally clustered, a strategy of selecting one idea from each cluster would be a reasonable approach to ensuring diversity. In the continuous manifold that characterizes mobility strategy ideation, cluster-based strategies are insufficient because any cluster boundary is somewhat arbitrary and the interior of each cluster may still contain many near-duplicate ideas. The weighted similarity gate addresses this by enforcing pair-wise minimum distance rather than cluster membership.

Patent Validation Results

The patent validation pipeline was tested on a multimodal emotion-sensing vehicle cabin idea drawn from the structured portfolio. The SerpAPI retrieval returned 38 candidate patents for the query constructed from the idea's core mechanism. The top embedding similarity score was 0.358, indicating that even the most similar patent found differed substantially from the idea in semantic content. The system assigned an Uniqueness score of 0.642, a Validity score of 1.0, and an overall patent worthiness score of 0.785, classifying the idea as Likely Patent-Worthy.

This result illustrates the value of embedding-based patent similarity over keyword-based approaches. A keyword search for "emotion sensing vehicle" would surface patents with similar surface vocabulary but potentially very different technical mechanisms. The embedding approach evaluates semantic content, which means patents that solve the same underlying problem through different vocabulary are correctly flagged as potentially relevant, while patents that share vocabulary but address different problems are correctly ranked lower. The 0.358 top similarity score represents a meaningful amount of conceptual distance from existing patents, which is the relevant measure for early-stage novelty assessment.

DISCUSSION

Why Invisible Enforcement Works

The single most important architectural insight from the experimental program is that diversity enforcement must be applied at the selection stage, not the generation stage. Version 1's failure in which 29% of ideas did not follow the 14-rule prompt framework despite explicit instructions illustrates a fundamental limitation of instruction-following as a diversity mechanism. Language models comply with constraints probabilistically; as the number and specificity of constraints increases, compliance rates fall because the model must simultaneously satisfy multiple objectives that may conflict within the training distribution.

By moving diversity enforcement into the selection mechanism, Kaleidoscope makes compliance deterministic rather than probabilistic. The gate either admits an idea or rejects it based on a computed similarity score; there is no probability of a near-duplicate slipping through because the system has no memory for instruction-following. The generation side produces candidates freely within the topic space, which means the generation model can focus on producing high-quality, coherent ideas rather than simultaneously managing constraint compliance. This separation of concerns is what allows Version 3 to achieve both zero duplicate pairs and improved novelty. It removes the trade-off between compliance and quality that plagued Version 1.

Dimension Weighting as Strategic Signal

The empirical derivation of dimension weights from unique value counts in the baseline library is both a methodological choice and a substantive finding. The fact that core mechanism has 200 unique values across 200 ideas while value type has only 15 unique values tells us something meaningful about how ideas in the mobility strategy domain differentiate themselves. Ideas that appear similar at the level of value type may be completely different at the level of mechanism. The opposite is less common: ideas that share a core mechanism tend to also share stakeholder, context, and value type, because the mechanism largely determines who benefits, where it applies, and what kind of value it creates.

This asymmetry justifies weighting mechanism most heavily and value type least heavily. A similarity gate that treats all dimensions equally would reject too many ideas that differ in mechanism but share a value type category, producing a portfolio that is overly sparse in some strategic directions and missing important convergent concepts that address similar needs through genuinely distinct approaches. The weighted formula preserves these convergent concepts while still preventing true duplication.

The Role of Radical Ideas in Portfolio Construction

One of the most striking results from the experimental program is the 28-fold increase in radical or transformative ideas in the structured portfolio relative to the unstructured baseline. This finding has implications beyond the immediate context of this project. It suggests that the scarcity of radical ideas in LLM-generated portfolios is not a fundamental limitation of language models but a consequence of the selection environment. In unstructured generation, radical ideas are rare because the model's probability distribution assigns them low probability relative to incremental ideas. When the selection mechanism explicitly rewards novelty and the 14-rule framework explicitly require a minimum share of radical ideas, the same models produce portfolios that are dominated by transformative concepts.

This has practical implications for how AI-assisted ideation should be designed. The question is whether the generation and selection architecture creates the conditions under which those ideas are produced and retained. Kaleidoscope creates those conditions but unconstrained prompting does not.

FUTURE STEPS

Through the process of building and validating Kaleidoscope, we encountered several limitations and open questions that point toward a clear agenda for future development. These are not abstract research directions, but specific gaps identified through working with the system across multiple runs and configurations.

Resolving Patent Coverage Limitations

The most significant limitation of the current implementation is the patent retrieval layer. Google Patents via SerpAPI provides broad coverage but not exhaustive coverage of the global patent landscape. In practice, this means the system may miss relevant prior art from jurisdictions or time periods that are underrepresented in Google's index, and it cannot perform the claims-level analysis that a full freedom-to-operate assessment requires. We encountered this limitation directly when validating the patent outputs where the similarity scores were directionally correct but could not be fully trusted as complete coverage of the relevant prior art.

The path forward is to build additional retrieval adapters for USPTO's PatentsView API, the European Patent Office's OPS API, and WIPO's PATENTSCOPE system, integrating them behind the same scoring interface as the current Google Patents path. Deduplication across sources and provenance tracking for each retrieved patent are necessary components of this expanded retrieval layer. This work was scoped for future development rather than included in the current implementation because it requires access to institutional API agreements and compliance review that were outside the project timeline.

Calibrating Thresholds Against Human Expert Judgment

The duplicate detection threshold of 0.70 and the k=5 novelty neighborhood size are governance parameters that were set based on empirical observation rather than calibration against human expert judgment. We observed that changing these parameters substantially changes every portfolio metric, moving the threshold from 0.70 to 0.60 nearly doubles the duplicate pair count for the same portfolio, while moving it to 0.80 reduces it to near-zero. The right threshold is the one at which the system's redundancy classifications agree with the judgments of domain experts who evaluate ideas based on strategic substance rather than surface similarity.

This calibration requires running structured studies in which Honda R&D researchers evaluate pairs of ideas from the baseline library and indicate whether they consider them strategically redundant, then comparing those human judgments to the cosine similarity scores for those pairs. The empirical threshold that maximizes agreement between human judgments and system classifications is the threshold that should be locked for production use. We recommend this calibration be conducted as a prerequisite to deploying Kaleidoscope in a live R&D workflow, as it will substantially increase trust in the system's outputs among stakeholders who need to rely on the duplicate pair counts and novelty labels as meaningful signals.

Domain-Tuned Embeddings

The current implementation uses text-embedding-3-small as a general-purpose embedding model. General-purpose models are trained on broad corpora and may not capture the fine-grained semantic distinctions that matter most in specialized domains. What counts as meaningfully different varies significantly by industry and context. In automotive engineering, the distinction between a sensor fusion architecture for urban driving and one for highway driving may be technically significant but appear as high similarity in a general embedding space because both concepts use similar vocabulary. In management consulting, two ideas that recommend different organizational restructuring approaches may appear semantically similar

because they share the same strategic vocabulary — transformation, efficiency, alignment — even when the underlying mechanisms and stakeholder implications are entirely distinct. In healthcare, ideas that apply the same intervention to different patient populations may be clinically significant distinctions that a general model collapses into near-identical embeddings. In financial services, two risk management frameworks that differ in their underlying assumptions about market behavior may score as highly similar because the language of risk, exposure, and mitigation is consistent across both. In energy, ideas that address grid stability through demand-side management versus supply-side generation may share enough vocabulary around load, capacity, and reliability to appear closer in embedding space than they are in engineering reality.

The framework is designed to be industry-agnostic at the architectural level, but the embedding layer is the component most sensitive to domain context. The dimension weights and similarity thresholds can be recalibrated through the empirical analysis described in this report for any new domain. However, a more durable solution is to use embedding models that have been exposed to domain-specific corpora, whether through fine-tuning on internal documents, retrieval-augmented generation (RAG) over a domain knowledge base, or selection of a specialized embedding model trained on industry literature. The appropriate approach depends on the resources and document access available in each deployment context. For organizations deploying Kaleidoscope in domains with large internal corpora of strategic documents, patents, or research reports, fine-tuning the embedding layer on that corpus is the most direct path to improving similarity precision for the distinctions that matter most in that domain.

Knowledge Graph Integration

The current output format, a four-tab Excel file, captures the ideas and their metadata but does not represent the relationships between ideas, between ideas and patents, or between ideas and Honda's strategic priorities in a form that supports complex queries. A knowledge graph representation in Neo4j would allow analysts to ask questions such as: which ideas share a core mechanism with existing Honda patents? Which ideas address the same stakeholder through different mechanisms? Which regulatory contexts appear most frequently across the highest-novelty ideas?

These queries are currently possible only through manual inspection of the Excel outputs, which does not scale to large portfolios or comparative analyses across multiple runs. A knowledge graph would make the portfolio navigable and quarriable in ways that complement rather than replace the current output format. We scoped this as future work because the implementation requires establishing a Neo4j deployment environment and designing the graph schema in collaboration with HRI stakeholders who would use it — tasks that require organizational coordination beyond what was feasible within the project timeline.

Human-in-the-Loop Weight Editing

The dimension weights are currently fixed at the empirically derived values from the baseline analysis. In practice, different strategic questions may warrant different weight configurations. A question focused on regulatory readiness might weight stakeholder and context more heavily to surface ideas targeted at specific regulatory actors and deployment environments. A question focused on technical differentiation might weight core mechanism even more heavily to maximize distance between the fundamental technical approaches in the portfolio.

We did not implement weight editing in the current UI because the appropriate interface for this feature requires domain expertise to design well. Presenting raw weight sliders to users who are not familiar with the embedding similarity methodology is likely to produce configurations that degrade rather than improve portfolio quality. The right interface would present the weight choices in terms of strategic intent and translate those choices into weight adjustments automatically.

CONCLUSION

Kaleidoscope delivers a domain-agnostic system for generating semantically diverse, non-redundant innovation portfolios with built-in proof of quality. The project began with Honda Research Institute's observation that LLM-generated idea portfolios look diverse but are not, and it ends with a validated framework that resolves this failure mode empirically: near-complete duplicate elimination while maintaining novelty on par with unconstrained generation, confirmed at p less than 0.000005.

The core contribution of this work is conceptual as much as technical. Treating ideation as a selection and optimization problem over semantic space rather than a generation problem changes what questions can be asked and answered about the quality of an idea portfolio. It makes redundancy measurable, novelty quantifiable, and the comparison between portfolios meaningful rather than subjective. These properties are what allow Kaleidoscope to function as decision-support infrastructure rather than just a brainstorming tool.

The experimental program demonstrated that the apparent trade-off between novelty and redundancy elimination is an artifact of operating in the domain saturation zone above 100 ideas, not a fundamental property of the structured approach. Within the pre-saturation zone, the framework achieves both objectives simultaneously. The 28-fold increase in radical ideas and the shift from 3% to 84.5% radical idea share in the portfolio demonstrate that the system does not merely reduce redundancy. It reshapes what kind of ideas make it into the portfolio, privileging transformative and cross-industry concepts that unstructured generation systematically underproduces.

For Honda Research Institute, the value is operational: a structured path from strategic question to actionable innovation portfolio, with transparent metrics that make the quality of that portfolio auditable and comparable across runs and topics. The longer-term value is institutional: a foundation for repeatable innovation intelligence infrastructure that accumulates knowledge about which conceptual territories have been explored, what the IP landscape looks like in each territory, and how portfolios evolve as the domain changes. These are the properties that distinguish a mature innovation process from one-off brainstorming, and they are what Kaleidoscope is designed to support.

As the Tepper School of Business framing puts it, the intelligent future is data-informed and human-driven. Kaleidoscope operationalizes this principle in the specific context of R&D ideation: the system measures, compares, and screens; the humans decide, prioritize, and act. The goal is never to replace the strategic judgment that researchers bring to the ideation process but to ensure that judgment is applied to a portfolio that deserves it: one where every idea represents a genuinely distinct direction worth evaluating.

REFERENCES

Anthropic. (2024). Claude API documentation. https://docs.anthropic.com

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. Scientific Reports, 13(1), 13601. https://doi.org/10.1038/s41598-023-40858-3

Li, L. (2024). Chain of Ideas: Revolutionizing research via novel idea development with LLM agents. arXiv preprint. https://doi.org/10.32388/aub766

OpenAI. (2024). OpenAI API documentation — Embeddings. https://platform.openai.com/docs/guides/embeddings

SerpAPI. (2024). Google Patents API documentation. https://serpapi.com/google-patents-api

Shen, J., Tenenholtz, N., Hall, J. B., Alvarez-Melis, D., & Fusi, N. (2024). Tag-LLM: Repurposing general-purpose LLMs for specialized domains. Proceedings of the 41st International Conference on Machine Learning (PMLR 235). https://arxiv.org/abs/2402.05140

World Intellectual Property Organization. (2024). World Intellectual Property Indicators 2024. WIPO. https://doi.org/10.34667/tind.50133

APPENDIX

Appendix A — Module Architecture

The Kaleidoscope codebase is organized into seven Python modules with well-defined responsibilities. agent_explorer.py implements the 14-rule diversity framework, the weighted similarity gate, batch generation, and sparse-cluster nudging. agent_refiner.py handles parallel concept brief expansion for the top N ideas by novelty rank. agent_patent.py manages SerpAPI patent retrieval, embedding similarity scoring, and GPT-based risk assessment. utils.py contains shared embedding functions, weighted cosine similarity computation, and pairwise novelty scoring. config.py holds all governance parameters — model names, dimension weights, thresholds, batch sizes — as a single source of truth. app.py implements the Gradio single-page UI with streaming progress, radar charts, patent gauges, and Excel download. agent.py provides a CLI entry point for headless batch runs and generalizability testing.

All seven modules are connected through a JSON idea object format that carries the full dimension label set and embedding vectors alongside the narrative idea text. This common format allows each agent to consume outputs from the previous stage without format conversion and makes the pipeline modular: any stage can be replaced or upgraded without changing the interface contract.

Appendix B — Evaluation Metrics

Weighted similarity is computed as the sum over all dimensions of the dimension weight multiplied by the cosine similarity of the two ideas' dimension embeddings. Weights sum to 1.0 and are defined in config.py as: core mechanism 0.30, problem opportunity 0.25, intended outcome 0.20, stakeholder 0.12, context 0.08, value type 0.05.

Novelty score is defined as one minus the mean of the top-5 cosine similarities to all other ideas in the portfolio on whole-idea embeddings. This is computed by the pairwise_novelty_scores function in utils.py. The k=5 neighborhood size was chosen to balance sensitivity to immediate nearest neighbors with stability across portfolio sizes.

Duplicate pairs are counted as the number of ideas whose nearest-neighbor cosine similarity on whole-idea embedding exceeds 0.70. This threshold was set based on empirical observation of the baseline library and represents a conservative definition of near-duplication — ideas below this threshold may still share thematic territory but are not considered strategically redundant under the current operational definition.

Novelty labels (High, Medium, Low) are assigned by the 75th and 25th percentile thresholds of the current portfolio's novelty score distribution. Ideas above the 75th percentile are labeled High, below the 25th percentile are labeled Low, and the remainder are labeled Medium. These thresholds are portfolio-relative, not absolute, which means the label distribution is always approximately 25% High, 50% Medium, and 25% Low for a normally distributed portfolio.

Appendix C — Data Provenance

All empirical claims in this report are traceable to documented source files. The baseline mean novelty figure was computed from the 200 idea library with embeddings spreadsheet using the k=5 novelty formula implemented in pairwise_novelty_scores in utils.py. Near-duplicate pair counts were computed from the same spreadsheet using nearest_neighbor.py. Clustering silhouette scores are documented in clustering_summary.xlsx generated by clustering.py. Archived Kaleidoscope run novelty scores and duplicate pair counts are drawn from the Diversity Snapshot tabs in the output Excel files stored in the outputs directory. Patent evaluation metrics are drawn from patent_worthiness_summary.csv produced by the research notebook run. Dimension weights are defined in config.py. Statistical test results were

computed from the Group A and Group B v3 novelty score distributions using scipy.stats.ttest_ind and scipy.stats.mannwhitneyu.

Appendix D — Dimension Label Extraction Prompt

The dimension labeling prompt instructs the model to extract exactly six fields for each idea: stakeholder, context, problem opportunity, core mechanism, intended outcome, and value type. Each field must not exceed eight words. The model is instructed to be consistent — reusing the same label when ideas address the same concept and avoiding synonyms unless the difference is meaningful. If an idea is ambiguous, the prompt instructs the model to choose the most plausible interpretation and to use broad rather than overly specific labels. The model is instructed to return only a JSON array with exactly one object per idea, with no additional text or markdown. Example output objects are provided to anchor the format and establish the expected level of specificity.

Appendix E — System Workflow