GAP DETECTION: MITIGATING HOMOGENIZATION THROUGH DEMAND AND DEFENSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models often produce semantically similar answers to open-ended questions, even when many plausible alternatives exist. Generating useful alternatives requires identifying both what the query calls for and what existing responses already cover. We introduce Gap Detection, a framework that searches for relevant but unoccupied directions in a shared text embedding space. Its objective, GapScore, combines query relevance (Demand) with dissimilarity to the nearest existing item (Defensibility). The framework optimizes this product on the unit sphere, maps candidate directions to concept words with a language model, and uses task-specific filtering and constrained prompting to turn those concepts into text. On five curated topics and 90 INFINITY-CHAT queries, the reported mean relative reductions in output Collision are 12.8% and 10.8%, respectively. We also instantiate the framework for generative engine optimization over four product markets and for iterative synthetic-data augmentation. In the latter setting, with 100 training examples per method, Gap Detection reduces Collision from 0.324 to 0.298 relative to undirected accumulation and increases the Vendi Score from 27.16 to 29.20; the accuracy difference is one example on a 40-example test set. Component ablations expose the trade-off between semantic separation and relevance, while also showing that better direction scores do not uniformly yield better downstream metrics. These results support explicit gap search as a practical mechanism for diversifying content, with task validity checked separately from geometric novelty.

1 INTRODUCTION

Open-ended questions invite multiple reasonable answers, yet language models frequently return variations of the same idea. A request for a metaphor about time, for example, can repeatedly elicit familiar images such as rivers or weaving. This pattern has been documented within and across models using INFINITY-CHAT, a collection of real-world open-ended queries (Jiang et al., 2025). Reference-free diversity metrics such as the Vendi Score likewise quantify the small effective number of semantic modes in such sets (Friedman & Dieng, 2023). For a user seeking alternative perspectives, lexical variation is insufficient: answers must differ in substance while continuing to address the question.

This problem suggests a distinction between sampling another answer and identifying an answer that is missing. Changing the decoding distribution can alter which tokens are sampled, but it does not explicitly represent the semantic content already covered by a collection of responses. Existing diversity-aware methods such as Maximal Marginal Relevance (MMR) trade query relevance against redundancy when reranking a fixed candidate set (Carbonell & Goldstein, 1998); learning-to-rank methods likewise score those candidates (Liu, 2009). Neither formulation searches unoccupied semantic space. Conversely, maximizing distance from previous responses alone can produce irrelevant concepts. A useful method must therefore keep new content connected to the user's demand while avoiding the nearest existing answer.

To address this problem, we introduce **Gap Detection**, a framework for discovering and realizing unoccupied semantic directions. Given a query and an occupied content set, the framework embeds both in a shared vector space (Mikolov et al., 2013), then searches for directions with high *GapScore*: the product of *Demand* (query relevance) and *Defensibility* (separation from the nearest occupied

item). Directions are selected sequentially and added to the occupied set, so later searches account explicitly for earlier coverage.

A high-scoring vector, however, is not yet a usable answer. Gap Detection bridges this gap through language-based realization. Embedding retrieval supplies a short list of words near a discovered direction; Qwen selects concepts that can answer the query, and a second filter checks their contextual plausibility. A text generator then uses these concepts while being prompted to avoid frequent concepts and sentence patterns in existing responses. Finally, the generated text is evaluated for novelty and repetition. The geometric search identifies where to explore, while the mapping and validation stages determine whether the proposed direction can be expressed as meaningful content.

The same relevance–coverage distinction appears beyond open-ended answering. In generative engine optimization (GEO), a candidate position should address user needs while differing from competitors' claims; GEO work shows that supplied content can affect which sources a generative engine cites (Aggarwal et al., 2024). Recursive training on generated data can reduce diversity (Shumailov et al., 2024; Alemohammad et al., 2024), whereas retaining real and synthetic data can alter that trajectory (Gerstgrasser et al., 2024). We study these as separate instantiations of the same objective, with setting-specific inputs, validity checks, and evaluation criteria.

Experiments examine both the effectiveness and the limits of this design. The reported mean relative Collision reduction is 12.8% over five curated topics and 10.8% over 90 open-ended queries. Additional experiments vary the source of the occupied responses and the model used to realize new directions. In four product markets, discovered positions receive citations in a controlled generative-engine simulation. In a small sentiment-data experiment, directed augmentation improves diversity relative to an accumulation baseline at the same final sample count. Ablations show that the product objective improves its intended relevance–separation score, but the full pipeline does not dominate every output-level metric.

Our contributions are threefold:

- **A coverage-aware objective.** We formulate gap discovery as joint optimization of Demand and Defensibility relative to an explicit occupied set.
- **A search-to-text pipeline.** We combine spherical optimization, language-model concept mapping, and task-specific validation so that a geometric direction is tested before it is accepted as content.
- **A cross-setting evaluation.** We evaluate the pipeline across open-ended generation, GEO positioning, and synthetic-data augmentation, reporting component trade-offs alongside aggregate gains. Appendix A provides the geometric existence result, conditional optimization results, and their proofs; the main text focuses on how the method operates and what the experiments establish.

2 Related Work

Output diversity and redundancy. Prior work shows that semantically repetitive outputs persist across model families and decoding configurations (Jiang et al., 2025). These findings motivate evaluating similarity at the level of complete responses. We use mean pairwise embedding similarity, termed Collision here, and complement it with the Vendi Score, a spectral measure of diversity (Friedman & Dieng, 2023). Gap Detection uses the occupied responses as an input to generation, rather than only as an evaluation set.

Relevance-aware diversification. Maximal Marginal Relevance (MMR) balances query relevance against similarity to previously selected documents (Carbonell & Goldstein, 1998). Gap Detection shares this redundancy-aware perspective. Its differences are a multiplicative objective, continuous search over embedding directions, and a realization stage that generates new text rather than only reordering existing documents. The product avoids an additive trade-off coefficient, but the complete pipeline still contains optimization and filtering hyperparameters. Prompted avoidance of repeated words is related to lexical constraints (Hokamp & Liu, 2017); unlike constrained decoding, our prompts do not guarantee that the generator obeys every constraint.

GEO and synthetic-data feedback. GEO studies how source content influences its visibility in generated answers (Aggarwal et al., 2024). We consider an earlier decision: which plausible position should new content express relative to existing claims? Separately, recursive synthetic-data training

can degrade generative distributions (Shumailov et al., 2024; Alemohammad et al., 2024), while retaining real and past synthetic data can mitigate collapse in studied settings (Gerstgrasser et al., 2024). Our augmentation experiment asks whether directing new samples toward semantic gaps adds value beyond retaining more samples. The documented procedure evaluates generated text with a downstream classifier and does not specify generative-model retraining, so it tests a narrower setting.

3 Gap Detection

3.1 Overview and Problem Setup

Gap Detection takes an occupied content set $\mathcal{C} = \{c_1, \ldots, c_N\}$ and a demand set $\mathcal{Q} = \{(q_j, f_j)\}_{j=1}^{M}$, where $f_j \geq 0$ and $\sum_j f_j = 1$. For open-ended answering, $\mathcal{C}$ contains existing responses and $\mathcal{Q}$ can contain a single query with weight one. Let $E(x) \in \mathbb{S}^{d-1}$ be a normalized text embedding. We write $\mathbf{e}_i = E(c_i)$ and $\mathbf{u}_j = E(q_j)$, so cosine similarity is the inner product.

The output is a collection of proposed directions and their textual realizations. The procedure has four stages: represent demand and occupied content; search for directions that balance relevance and separation; map the directions to plausible concepts; and generate and check text. These stages address different failure modes. Search discourages repetition, concept mapping makes vectors interpretable, and validation rejects directions or outputs that do not satisfy the task. We describe the shared mechanism first and then specify its domain adaptations.

Figure 1 summarizes the search-to-text interface. The occupied set and query enter the same embedding space, while only the direction is passed to the concept stage through retrieved candidate words. The verifier closes the loop: it checks coherence, task relevance, and novelty, then either returns an accepted output or feeds the rejection back into the next direction search. This separation makes it possible to diagnose whether a failure comes from geometric search, concept mapping, or realization.

3.2 Scoring Relevant and Unoccupied Directions

Definition 1 (Demand)**.** *For a unit direction* $\mathbf{z}$*, the frequency-weighted relevance to the demand set is*

$$\mathrm{Demand}(\mathbf{z}) = \sum_{j=1}^{M} f_j \langle \mathbf{u}_j, \mathbf{z} \rangle = \langle \mathbf{v}, \mathbf{z} \rangle, \qquad \mathbf{v} = \sum_{j=1}^{M} f_j \mathbf{u}_j. \tag{1}$$

Demand preserves the connection to the task. For one query it is simply query–direction cosine similarity; for multiple queries it weights their contributions. These weights represent the supplied demand specification, not a learned relevance model.

Definition 2 (Defensibility)**.** *For an occupied embedding set* $\mathcal{O}$*, define*

$$\mathrm{Def}(\mathbf{z}; \mathcal{O}) = 1 - \max_{\mathbf{e} \in \mathcal{O}} \langle \mathbf{e}, \mathbf{z} \rangle. \tag{2}$$

The nearest occupied item determines this score. A candidate close to even one existing item is penalized, regardless of its distance from the others. Defensibility lies in $[0, 2]$ for unit vectors; it can exceed one when every occupied item has negative similarity to the direction. It measures geometric separation, not factual validity or commercial defensibility.

Definition 3 (GapScore)**.** *The joint direction score is*

$$\mathrm{GapScore}(\mathbf{z}; \mathcal{O}) = \mathrm{Demand}(\mathbf{z}) \, \mathrm{Def}(\mathbf{z}; \mathcal{O}). \tag{3}$$

For nonnegative Demand, the product rewards both relevance and separation and vanishes when either is zero. Directions with negative Demand receive nonpositive scores. Thus, an empty region is not automatically desirable: it must also align with demand. The product can still trade off two positive factors, so it supplies a preference rather than a hard lower bound on either factor. Appendix A.3 discusses its connection to bargaining objectives under explicit assumptions.

Demand-only optimization ignores the occupied set, whereas Defensibility-only optimization can leave the query. Their product couples relevance and separation; concept and output checks then supply the task-specific validity test.

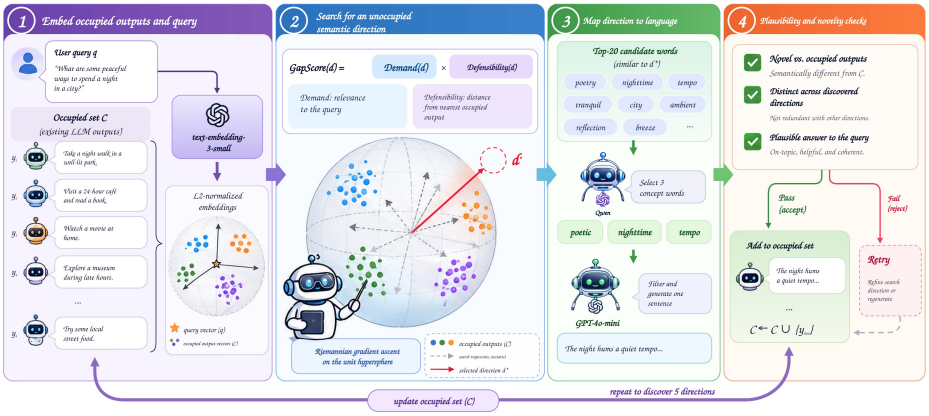

Figure 1: Overview of Gap Detection. The pipeline embeds existing content and the query, searches the unit hypersphere for a high-GapScore direction, maps that direction to a concrete concept, and verifies the resulting candidate before accepting it. A rejected candidate triggers another search; an accepted direction is added to the occupied set so later directions cover a different region.

3.3 Searching for Multiple Directions

We approximately maximize Eq. (3) on the unit sphere. Let $\mathbf{e}_*$ be a nearest occupied vector at the current iterate. Where that nearest vector is unique, the Euclidean gradient is

$$\mathbf{a} = \mathrm{Def}(\mathbf{z}; \mathcal{O})\mathbf{v} - \mathrm{Demand}(\mathbf{z})\mathbf{e}_*. \tag{4}$$

The first term moves toward demand, and the second moves away from the closest occupied direction. We project this update onto the sphere's tangent space and normalize after stepping:

$$\mathbf{g} = (I - \mathbf{z}\mathbf{z}^\top)\mathbf{a}, \qquad \mathbf{z} \leftarrow \frac{\mathbf{z} + \eta\mathbf{g}}{\|\mathbf{z} + \eta\mathbf{g}\|}. \tag{5}$$

This is a tangent update followed by a normalization retraction (Absil et al., 2008). At a nearest-neighbor tie, an active neighbor supplies the update direction. Multiple starts reduce reliance on any one initialization; they do not certify global optimality for this nonsmooth objective.

To obtain K directions, the search at round k uses

$$\mathcal{O}_k = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \cup \{\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}\}. \tag{6}$$

Adding a selected direction makes exact rediscovery have zero Defensibility at the next round. The procedure therefore encourages successive directions to explore different regions while retaining the same demand vector. The reported implementation uses demand-aligned, perturbed-demand, and random starts, 15 restarts per strategy, 100 updates per restart, and $\eta = 0.05$. The answering and GEO experiments use $K = 5$.

The geometry also provides a useful diagnostic. If demand has a nonzero component orthogonal to the span of occupied content, normalizing that component gives a direction with Defensibility exactly one and positive Demand. This yields a lower bound on the best achievable geometric score. It does not guarantee that a corresponding natural-language answer exists. The precise condition, proof, and numerical comparison are in Appendix A.1; the language stages below remain necessary even when the bound holds.

3.4 Mapping Directions to Plausible Concepts

A direction must be translated into concepts before a generator can use it. We construct a vocabulary from the Qwen2.5-3B-Instruct tokenizer, filter it to 13,409 English words, and embed these words

with the same encoder used for the content. A query- and corpus-dependent screening score retains 10,000 words. For each direction, cosine retrieval then selects 20 unused candidate words near that direction. This screening is computational; it does not decide whether a word answers the query.

Qwen receives the query and the candidate words and selects up to three concepts that could form concrete answers. GPT-4o-mini subsequently checks whether the selected concepts fit the requested kind of answer, including whether a proposed word merely restates the query. If fewer than three words pass, the pipeline examines further candidates, with at most ten rounds. Accepted words are not reused across directions. Appendix B gives the vocabulary score and candidate-handling details, and Appendix F gives the available prompt templates.

This separation matters because geometric proximity and linguistic usefulness are different tests. A nearby token may be a synonym of the question rather than an answer to it; a highly separated token may be unrelated to the task. The language-model stages evaluate these distinctions after the embedding search. They provide a plausibility screen, not a correctness certificate.

3.5 Generating and Checking Text

For each accepted concept group, GPT-4o-mini generates one sentence conditioned on the original query. The prompt also includes two kinds of avoidance instructions derived from the occupied content. First, the pipeline identifies 12 frequent words close to the occupied-content centroid and asks the generator to avoid them. Second, it extracts recurring sentence patterns and asks for a different construction. Concept guidance specifies what to express, while these instructions discourage a return to the content and phrasing already represented in $\mathcal{C}$.

The realized sentences are embedded again. A novelty check requires the maximum similarity to the original occupied responses to be below 0.7. A separate check compares opening words, and another checks whether the maximum pairwise similarity among the new sentences is below 0.8. The opening-word check is lexical; the two similarity checks are semantic proxies. Neither establishes human-perceived creativity or factual accuracy. Measuring the realized text is necessary because a generator may move away from the direction suggested by its concept words.

The concept words form an inspectable intermediate representation: a failure can be attributed to direction search, concept selection, or sentence realization.

3.6 Adapting the Framework to GEO and Synthetic Data

GEO positioning. The occupied set becomes a collection of competitor product claims, and demand consists of weighted search queries. The concept filter asks whether a candidate is plausible within the product category; the generator turns accepted words into a proposed claim. A controlled citation simulation then measures its visibility. Candidate claims are ranked by direction GapScore multiplied by the visibility score. Because the framework does not receive the target product's specifications, its output is a proposed market position requiring factual verification before use.

Synthetic-data augmentation. The occupied set becomes the current synthetic sample pool, and task-class descriptions with class-frequency weights specify demand. The concept filter checks task relevance, and the generator creates labeled examples. For sentiment classification, VADER (Hutto & Gilbert, 2014) supplies an additional polarity check: compound scores above 0.05 pass as positive and below -0.05 as negative. Candidate words are also reranked by absolute polarity. Directed methods retry generation up to four rounds to reach the accepted-example target. Final comparisons use $K = 1$ per label and generation; a separate direction-set analysis uses $K = 3$. The downstream classifier is trained on the resulting data and evaluated on a disjoint real test set. Its accuracy is an outcome measure, not an input to direction selection.

4 Experiments

We ask whether Gap Detection reduces repetition in realized outputs, how each score component affects the result, and whether the search mechanism remains useful when the occupied content changes from responses to claims or training examples. The three settings share the embedding and search machinery but use separate task-specific evaluation protocols.

4.1 Evaluation Protocol and Metrics

All text embeddings use `text-embedding-3-small` with 1,536 dimensions and L2 normalization. Qwen2.5-3B-Instruct-Q6_K supplies concept decoding; GPT-4o-mini supplies the default plausibility filtering and text generation. Unless explicitly varied, these components are fixed across the direction-selection ablations.

Definition 4 (Collision and novelty)**.** *For an embedded text set $X = \{x_i\}_{i=1}^n$, $n \geq 2$, and a generated text g, define*

$$\text{Collision}(X) = \frac{2}{n(n-1)} \sum_{i<j} \langle E(x_i), E(x_j) \rangle \,, \tag{7}$$

$$\text{Novelty}(g; \mathcal{C}) = 1 - \max_{c \in \mathcal{C}} \langle E(g), E(c) \rangle \,. \tag{8}$$

Lower Collision indicates lower mean similarity; a value of zero means zero average similarity, not necessarily pairwise orthogonality. Novelty measures separation of a realized text from its nearest original reference. We report relative Collision reduction as $100(C_{\text{before}} - C_{\text{after}})/C_{\text{before}}$ and preserve the reported mean of these relative reductions. A mean relative reduction need not equal the reduction computed from two aggregate means. The 90-query novelty statistic is the fraction of queries whose best generated sentence has Novelty greater than 0.3.

For synthetic datasets, we also use the Vendi Score (Friedman & Dieng, 2023), $\text{VS} = \exp(-\sum_i \lambda_i \log \lambda_i)$, where λ_i are eigenvalues of the normalized cosine Gram matrix. It captures spectral diversity beyond a mean similarity. Downstream utility is measured by classification accuracy. These metrics are distinct from direction GapScore: optimizing an embedding objective is not itself evidence of improved output quality. The available results are point estimates; no repeated-run uncertainty estimates are supplied.

4.2 Open-Ended Generation

Setup. The topic study uses five manually selected themes, with five query variants per theme and 20 GPT-4o responses as the occupied set. A second study uses 90 manually selected INFINITY-CHAT queries. Five directions and corresponding sentences are produced for each query. Table 1 summarizes the before/after Collision measurements. Detailed query-group results and reporting qualifications are given in Appendix C.

Table 1: Reported Collision before and after Gap Detection. The last column is the reported mean relative reduction. Topic averages and the 90-query aggregate are separate evaluations.

Evaluation	Before ↓	After ↓	Reduction ↑
Time metaphor	0.7362	0.6573	10.7%
Loneliness metaphor	0.7449	0.6634	10.9%
Redefine courage	0.8046	0.7160	11.0%
Memory as substance	0.8163	0.7080	13.3%
Hope via natural elements	0.8178	0.6712	17.9%
Five-topic mean (reported)	0.784	0.683	12.8%
90 INFINITY-CHAT queries	0.7766	0.6915	10.8%

Results. All five topics exhibit lower reported Collision, with reductions from 10.7% to 17.9%. The 90-query experiment shows a similar aggregate pattern, with Collision decreasing from 0.7766 to 0.6915 and a largest single-query reduction of 20.5%. The best sentence passes the novelty threshold on 88 of 90 queries (97.8%). These observations indicate separation from existing responses under the chosen encoder; they do not establish that the sentences are better answers according to human judgments.

The study does not include a matched rerun of temperature or min-p baselines. The decoding experiments in prior work (Jiang et al., 2025) provide motivation, but their results are not a controlled numerical baseline for this pipeline. Likewise, direction search, concept filtering, and avoidance

prompting contribute jointly to the observed end-to-end change. The ablation below isolates the direction-selection rule within that shared pipeline.

Component ablation. We compare random directions, Demand-only search, Defensibility-only search, and the product objective on five topics (Table 2). GapScore achieves the highest direction score, 0.375, compared with 0.298 for Demand-only and -0.228 for Defensibility-only. However, Defensibility-only achieves the largest Collision reduction, and random directions achieve the highest novelty-check pass rate. The full method's pass rate is 84% in this separate ablation. Thus, both terms help balance the geometric objective, but the ablation does not establish uniform superiority in realized-text metrics. The relatively strong random-direction result also suggests that the shared filtering and prompting stages make a substantial contribution.

Table 2: Direction-selection ablation on five topics. Direction GapScore evaluates the optimized vectors; the other columns evaluate realized outputs. Best values in each column are bold. These runs are separate from Table 1.

Method	Direction GapScore ↑	Collision reduction ↑	Novelty pass ↑
Random direction	0.003	10.3%	**100%**
Demand only	0.298	7.8%	92%
Defensibility only	-0.228	**11.5%**	96%
GapScore	**0.375**	10.7%	84%

4.3 Transfer across Content Sources and Generators

The occupied responses and the realization model affect different stages of Gap Detection. Changing the source of existing responses changes which directions the search must avoid. Changing the realization model instead tests how effectively a discovered direction is turned into text. We evaluate these dependencies separately on three topics.

Changing the occupied set. Each of four models supplies its own set of 20 occupied responses, while GPT-4o-mini remains the realization model. Collision decreases for every source (Table 11), with reported reductions of 8.8–12.7% and a mean of 10.4%. The novelty-pass rate averages 96.7%. This result supports using the same search and realization procedure across the tested content sources; it does not require the occupied responses to come from the generator that will produce the new answer.

The separate-source study reports reductions of 11.2% (GPT-4o-mini), 12.7% (Claude-Sonnet-4.5), 8.8% (Gemini-2.0-Flash), and 8.8% (Qwen2.5-3B), averaging 10.4% with a 96.7% novelty-pass rate. The complete table is in Appendix C.2.

A mixed-source experiment combines five responses from each of these four models into one occupied set. Its recorded Collision values are 0.602 before intervention and 0.561 afterward, with a reported reduction of 7.4%. This tests whether the search can account for content contributed by several models simultaneously; supplementary source-similarity diagnostics are in Appendix C.2.

Changing the realization model. With 20 GPT-4o-mini responses held fixed as the occupied set, the reported Collision reduction is 9.0% for GPT-4o-mini, 11.6% for GPT-4o, and 14.8% for Claude-Sonnet-4.5, with novelty-pass rates of 80.0%, 100.0%, and 100.0%. The search can therefore be robust to the source set while remaining sensitive to how a direction is expressed. These output comparisons do not establish a general ranking of model quality.

4.4 GEO Market Positioning

Setup and visibility measurement. We use Amazon product metadata (Hou et al., 2024) for water filters, protein powder, sports water bottles, and yoga mats. Grouping claims by brand and filtering sparse brands leaves 3–17 brands and 12–142 claims per market. Each market has 15 manually designed, weighted search queries. The method searches for five directions and realizes them as proposed claims.

To evaluate a claim, GPT-4o-mini receives four competitor claims as sources [1–4] and the candidate as source [5], then answers a market query with inline citations. We compute a position-adjusted cited-word share inspired by GEO (Aggarwal et al., 2024):

$$\mathrm{PAWC}(g) = \frac{\sum_{j=0}^{m-1} \mathbb{1}\{[5] \in s_j\}\, |s_j| \exp(-j/m)}{\sum_{j=0}^{m-1} |s_j|}, \tag{9}$$

where s_j is an answer sentence and $|s_j|$ its word count. This is the implementation recorded for these experiments, not a measurement of live retrieval or a calibrated citation probability. The same simulation supplies the visibility factor used to select the best candidate, so best-candidate results are selection-set results.

Table 3: GEO results. Top: best direction per market, selected using GapScore × PAWC. Bottom: a separate ablation averaged across four markets. The blocks have different aggregation rules.

Market or method	Direction GapScore ↑	PAWC ↑	Product ↑
Water filters	0.410	0.518	0.213
Protein powder	0.479	0.540	0.258
Sports water bottles	0.398	0.513	0.204
Yoga mats	0.403	0.524	0.211
Reported market mean	0.423	0.524	0.221
Random direction	−0.006	0.469	−0.003
Demand only	0.279	**0.508**	0.143
Defensibility only	−0.164	0.487	−0.080
GapScore	**0.307**	0.482	**0.149**

Results and interpretation. The selected positions have a reported mean product score of 0.221 (Table 3). For water filters, the words *campground, wilderness, hike* express an outdoor-use direction; its GapScore is 0.410 and PAWC is 0.518. A more generic direction, *screened, cleansing, reservoir*, has a higher PAWC of 0.544 but a lower GapScore of 0.135, giving a smaller product of 0.073. This example illustrates the intended distinction between being cited and identifying a differentiated position.

In the separate ablation, GapScore has the highest product score, 0.149 versus 0.143 for Demand-only, while Demand-only has the highest PAWC, 0.508 versus 0.482. The evidence therefore supports better joint relevance–separation scoring in this simulation, rather than a standalone visibility advantage. Complete market factors, concept words, and water-filter directions are in Appendix D.

The factors expose the market trade-off: protein powder has Demand 0.383 and Defensibility 1.249, while yoga mats have Demand 0.763 and Defensibility 0.528. Defensibility above one does not verify a product property.

4.5 Iterative Synthetic-Data Augmentation

Setup. We use SST-2 sentiment examples (Socher et al., 2013), with ten seed examples per class and a disjoint test slice of twenty examples per class. Across five generation rounds, accumulation methods target eight additional accepted examples per class per round. A TF-IDF logistic-regression classifier is trained on each round's data. The final comparison has 100 training examples per method. No Intervention replaces the preceding pool with newly generated text; Data Accumulation retains prior data and adds undirected samples; GapScore retains prior data and adds directed samples with the filtering in Section 3.6. These are controlled baselines inspired by replacement and accumulation studies, not exact replications of their model-training protocols.

Diversity and downstream utility. GapScore lowers Collision from 0.324 to 0.298 relative to Data Accumulation and raises Vendi from 27.16 to 29.20 (Table 4). Its 62.5% accuracy exceeds the two baselines' 60.0% by one test example. This is a positive point estimate in a small task, not evidence of a statistically reliable accuracy gain. Matching accepted sample counts also does not match API-call budgets: directed generation may make additional calls when candidates fail the polarity check.

Table 4: Synthetic-data results at generation five. All three methods have 100 training examples. Accuracy is evaluated on 40 disjoint real examples; the 2.5-point difference corresponds to one prediction.

Method	Collision ↓	Vendi Score ↑	Accuracy ↑
No Intervention	0.362	9.06	60.0%
Data Accumulation	0.324	27.16	60.0%
GapScore	**0.298**	**29.20**	**62.5%**

Separate four-generator runs show lower Collision for GapScore than accumulation in all four cases, with differences between -0.019 and -0.028. Vendi improves in three cases, but decreases by 0.74 for Qwen2.5-3B. These reruns have different numerical deltas from Table 4; they are reported separately in Appendix E. No model-wise downstream accuracy table is available for these reruns. The results support a recurring reduction in average similarity, while exposing a case where spectral diversity moves in the opposite direction.

Direction-set analysis. A separate $K = 3$ analysis measures how a set of directions covers demand. For $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, define

$$\mathrm{SetGapScore}(S) = \left(\sum_j f_j \max_{\mathbf{z} \in S} \langle \mathbf{u}_j, \mathbf{z} \rangle \right) \frac{1}{|S|} \sum_{\mathbf{z} \in S} \mathrm{Def}(\mathbf{z}; \mathcal{O}_1). \tag{10}$$

GapScore obtains 0.276, compared with 0.199 for Demand-only, 0.023 for random directions, and -0.029 for Defensibility-only. Demand-only repeats the same direction (pairwise cosine 1.000), whereas GapScore's selected directions have similarities of 0.11–0.39. The maximum in Eq. (10) does not reward repeated coverage of the same query. This is a diagnostic of the intended direction objective, not an independent downstream quality metric or a submodular guarantee for the full product. The $K = 3$ results should not be combined with the $K = 1$ augmentation comparison.

5 Discussion and Limitations

The central result is that an explicit representation of prior content can guide a language pipeline toward less repetitive outputs. The method separates three questions: where the occupied set leaves room, whether a concept fits the task, and whether its realization is useful. The ablations show why those questions should remain separate. Better geometric scores can coexist with weaker novelty-check rates or lower citation share; direction optimization alone does not explain all end-to-end gains.

The evaluation has several limits. Collision and Vendi depend on the encoder used for both search and evaluation, and no independent human assessment of answer quality is reported. The query subset is manually selected. Some aggregation details remain unresolved in the experimental record, and raw outputs and repeated-run uncertainty are needed for a complete replication. GEO results use fixed source positions in a simulation, do not test live retrieval, and do not verify that a proposed claim is true of a target product. Sparse markets can also make apparent gaps easier to find. Finally, the sentiment study uses a small seed and test set, a lightweight classifier, and no documented generator-retraining procedure; it does not establish prevention of collapse in recursively retrained foundation models. VADER filtering may favor simple polarity cues, and matched training-set sizes do not isolate its effect from direction selection.

6 Conclusion

Gap Detection makes prior semantic coverage an explicit input to generation. It combines Demand and Defensibility to search for directions, maps them to concepts, and uses task-specific checks to realize them as text. The reported experiments show lower Collision across the open-ended generation studies and useful diversity gains in a small augmentation task, with a GEO study illustrating differentiated position discovery. The mixed ablations also identify the remaining challenge: translating a well-separated, relevant direction into content whose quality improves under independent evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for several assistive purposes. We used them to run parts of the experimental pipeline (executing the embedding, search, mapping, and verification steps of the method and collecting the resulting metrics), to retrieve and discover related work (identifying and locating candidate citations for concepts and comparisons already determined by the authors), to draft and polish sections of the manuscript's writing (producing initial phrasing for some passages and improving clarity, wording, and consistency of the prose), and to check the derivations underlying the propositions and theorem in Appendix A against standard results after the authors formulated the claims. We have not used generative AI tools to design the core method or decide which results to report; those decisions were made by the authors, and no other disclosure items from the ICLR AI Policy apply to this work. We have reviewed all AI-assisted work: we checked that AI-run experiments used the intended settings and reproduced consistent numbers across repeated runs, we verified that every AI-suggested citation is accurate and genuinely relevant to the point it supports, we independently re-derived the proof steps that AI helped check, and we verified that the drafted and polished prose preserved the original meaning without introducing unsupported claims. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The reported evaluations use existing datasets and controlled generated content. The GEO experiment evaluates proposed claims in a citation simulation; a plausibility filter does not establish that a real product has the proposed properties. Any deployment would require factual verification. Similarly, semantic novelty does not imply that an answer is safe, correct, or useful. Dataset licenses, privacy requirements, and provider terms must be respected when distributing the data and generated examples. The sentiment experiment trains a small downstream classifier and should not be interpreted as validation of a deployed model.

REPRODUCIBILITY STATEMENT

Section 3 specifies the scoring and optimization procedure. Appendix B records implementation settings, Appendix F reproduces the available prompts, and Appendix A contains the theoretical statements and proofs. Appendices C–E give the supplementary results. The materials contain aggregate results rather than executable experiment code, raw outputs, and a complete run manifest.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. *GEO: Generative Engine Optimization*. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. arXiv:2311.09735.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. *Self-Consuming Generative Models Go MAD*. *ICLR 2024*. arXiv:2307.01850.

Anthropic. *System Card: Claude Sonnet 4.5*. `https://www.anthropic.com/claude-sonnet-4-5-system-card`, 2025.

Allan Borodin, Hyun Chul Lee, and Yuli Ye. *Max-Sum Diversification, Monotone Submodular Functions and Dynamic Updates*. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pp. 155–166, 2012.

Nicolas Boumal, P.-A. Absil, and Coralia Cartis. *Global Rates of Convergence for Nonconvex Optimization on Manifolds*. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Jaime Carbonell and Jade Goldstein. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, 1983.

Dan Friedman and Adji Bousso Dieng. *The Vendi Score: A Diversity Evaluation Metric for Machine Learning*. *Transactions on Machine Learning Research*, 2023. arXiv:2210.02410.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data*. *First Conference on Language Modeling (COLM)*, 2024. arXiv:2404.01413.

Google DeepMind. *Gemini 2.0 Flash Model Card*. `https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf`, 2025.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. *The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text*. arXiv:2311.09807, 2023.

Chris Hokamp and Qun Liu. *Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1535–1546, 2017.

Seyedehsomayeh Hosseini and André Uschmajew. *A Riemannian Gradient Sampling Algorithm for Nonsmooth Optimization on Manifolds*. *SIAM Journal on Optimization*, 27(1):173–189, 2017.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. *Bridging Language and Items for Retrieval and Recommendation*. arXiv:2403.03952, 2024.

C.J. Hutto and Eric Gilbert. *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. *Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)*. NeurIPS 2025 Best Paper. arXiv:2510.22954.

W. Chan Kim and Renée Mauborgne. *Blue Ocean Strategy: How to Create Uncontested Market Space and Make the Competition Irrelevant*. Harvard Business School Press, 2005.

Weiyue Li, Mingxiao Song, Zhenda Shen, Dachuan Zhao, Yunfan Long, Yi Li, Yongce Li, Ruyi Yang, and Mengyu Wang. *LLM Review: Enhancing Creative Writing via Blind Peer Review Feedback*. arXiv:2601.08003, 2026.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781, 2013.

John F. Nash. *The Bargaining Problem*. *Econometrica*, 18(2):155–162, 1950.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. *An Analysis of Approximations for Maximizing Submodular Set Functions—I*. *Mathematical Programming*, 14(1):265–294, 1978.

OpenAI. *GPT-4o System Card*. `https://openai.com/index/gpt-4o-system-card/`, 2024.

Qwen Team, Alibaba Group. *Qwen2.5 Technical Report*. arXiv:2412.15115, 2024.

A. H. G. Rinnooy Kan and G. T. Timmer. *Stochastic Global Optimization Methods, Part II: Multi-Level Methods*. *Mathematical Programming*, 39(1):57–78, 1987.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. *Okapi at TREC-3*. In *TREC-3 Proceedings*, pp. 109–126, 1994.

Ali Safari and Sahar Babaei. *Research Novelty in Information Systems Journals After ChatGPT: Differences Across Institutional Language Contexts*. arXiv:2603.22510, 2026.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Style Transfer from Non-Parallel Text by Cross-Alignment*. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. *The Curse of Recursion: Training on Generated Data Makes Models Forget*. arXiv:2305.17493, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. *AI Models Collapse When Trained on Recursively Generated Data*. *Nature*, 631(8022):755–759, 2024.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

A Theoretical Analysis

This appendix collects all formal results and proofs. The geometric statements concern vectors in an embedding space. They do not establish that every such vector can be realized as valid text. Results borrowed from smooth optimization and submodular maximization are stated with their additional assumptions, rather than applied unconditionally to the implemented search.

A.1 Existence of a Geometric Gap

Let $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{S}^{d-1}$, let $\mathcal{S} = \operatorname{span}(\mathcal{E})$, and let P and $P^\perp = I - P$ be the orthogonal projectors onto $\mathcal{S}$ and its complement. The demand vector is $\mathbf{v} = \sum_j f_j \mathbf{u}_j$.

Lemma A.1 (Projection invariance)**.** *For any vector* $\mathbf{z}$*, the inner-product extension of Defensibility satisfies*

$$\operatorname{Def}(\mathbf{z}; \mathcal{E}) = 1 - \max_i \langle \mathbf{e}_i, P\mathbf{z} \rangle .$$

Proof. Each $\mathbf{e}_i$ lies in $\mathcal{S}$, so $P\mathbf{e}_i = \mathbf{e}_i$. Self-adjointness gives $\langle \mathbf{e}_i, \mathbf{z} \rangle = \langle P\mathbf{e}_i, \mathbf{z} \rangle = \langle \mathbf{e}_i, P\mathbf{z} \rangle$. Taking the maximum proves the identity. □

Theorem A.2 (Unmet-demand lower bound)**.** *Suppose* $a = \left\| P^\perp \mathbf{v} \right\| > 0$*. Then the unit vector*

$$\mathbf{z}_0 = \frac{P^\perp \mathbf{v}}{\|P^\perp \mathbf{v}\|}$$

satisfies $\operatorname{Def}(\mathbf{z}_0; \mathcal{E}) = 1$ *and* $\operatorname{Demand}(\mathbf{z}_0) = a$*. Consequently,*

$$\max_{\mathbf{z} \in \mathbb{S}^{d-1}} \operatorname{GapScore}(\mathbf{z}; \mathcal{E}) \geq a. \tag{11}$$

Proof. For every occupied vector, $\left\langle \mathbf{e}_i, P^\perp \mathbf{v} \right\rangle = 0$ by orthogonality. Therefore the maximum occupied similarity to $\mathbf{z}_0$ is zero, yielding Defensibility one. Also,

$$\langle \mathbf{v}, \mathbf{z}_0 \rangle = \frac{\left\langle \mathbf{v}, P^\perp \mathbf{v} \right\rangle}{a} = \frac{\left\| P^\perp \mathbf{v} \right\|^2}{a} = a.$$

Multiplying gives $\operatorname{GapScore}(\mathbf{z}_0; \mathcal{E}) = a$. Since $\mathbf{z}_0$ is feasible, the maximum cannot be smaller. □

Corollary A.3 (Dimension of the orthogonal complement)**.** *If* $r = \dim \mathcal{S} < d$*, then* $\mathcal{S}^\perp$ *has dimension* $d - r$*, and every unit vector in it has Defensibility one.*

Proof. The dimension statement is the orthogonal decomposition of $\mathbb{R}^d$. For a unit vector in $\mathcal{S}^\perp$, every occupied inner product is zero, giving the score directly. □

For $N = 20$ and $d = 1536$, the complement has dimension at least 1516. This dimension count holds for any collection of 20 vectors in this space; it is not itself evidence of semantic homogenization. Positive-demand existence further requires $P^\perp \mathbf{v} \neq 0$, and linguistic realizability remains a separate question. Defensibility one is not its maximum possible value, since negative occupied similarities can give values above one.

Proposition A.4 (Attainment of the maximum)**.** *For a finite, nonempty occupied set, GapScore attains a maximum on* $\mathbb{S}^{d-1}$.

Proof. Demand is linear and Defensibility is one minus the maximum of finitely many continuous functions. Their product is continuous. The sphere is compact, so the extreme-value theorem applies. □

A.2 A One-Sided Perturbation of the Orthogonal Direction

Proposition A.5 (Directional improvement condition)**.** *Assume* $a = \left\|P^{\perp}\mathbf{v}\right\| > 0$ *and* $b = \|P\mathbf{v}\| > 0$. *Put* $\mathbf{u} = P\mathbf{v}/b$ *and* $m = \max_i \langle \mathbf{e}_i, \mathbf{u}\rangle$. *Along* $\mathbf{z}(\theta) = \cos\theta\, \mathbf{z}_0 + \sin\theta\, \mathbf{u}$, *the right derivative at zero is*

$$\frac{d}{d\theta}\,\mathrm{GapScore}(\mathbf{z}(\theta); \mathcal{E})\bigg|_{0^+} = b - am.$$

In particular, $b > am$ *is sufficient for a local improvement along this positive perturbation.*

Proof. Orthogonality of $\mathbf{z}_0$ and $\mathbf{u}$ makes $\mathbf{z}(\theta)$ a unit vector. Its Demand is $a\cos\theta + b\sin\theta$. For $\theta \geq 0$ sufficiently small, $\sin\theta \geq 0$, so the maximum occupied inner product is $m\sin\theta$. Thus the product equals $(a\cos\theta + b\sin\theta)(1 - m\sin\theta)$ on that side. Differentiation at 0^+ yields $b - am$. □

The statement is one-sided because the active maximum can change on reversing the perturbation. A nonpositive derivative along this path does not prove global tightness of Eq. (11), nor optimality against all tangent directions.

Table 5: Recorded numerical comparison for the geometric lower bound. The search column is the best observed score, not a certified global optimum. All five occupied matrices are reported to have rank 20, giving $d - r = 1516$.

Topic	Lower bound a	Search score	Difference	$b = \|P\mathbf{v}\|$
Time metaphor	0.6438	0.6492	0.0054	0.5841
Loneliness metaphor	0.7474	0.7493	0.0019	0.6644
Redefine courage	0.7156	0.7389	0.0233	0.6985
Memory as substance	0.5455	0.6011	0.0556	0.8381
Hope via natural elements	0.7731	0.7893	0.0162	0.6343

The report records orthogonality residuals of order 10^{-16} for these calculations. The observed search values exceed the lower bounds by approximately 0.25–10.2% relative to the bounds. These are empirical comparisons to returned solutions, not approximation ratios to the unknown optima. The supplied values of b/a are 0.907, 0.889, 0.976, 1.536, and 0.820 in table order; the observed differences are not monotone in that ratio. The perturbation also depends on m, which the table does not report.

The orthogonal direction can be constructed using an orthonormal basis U of the occupied span, with residual $\mathbf{v} - UU^{\top}\mathbf{v}$. It is a feasible reference point whenever this residual is nonzero. The same construction remains applicable after adding previously selected directions, provided the new residual is nonzero. The experiments report comparisons to this construction, but do not establish that it was used as an additional initializer in every main run.

A.3 What the Product Objective Implies

The product has two elementary properties on the positive region: a zero factor gives zero score, and multiplying either factor by a positive constant preserves the ranking of candidates. These facts motivate its use without claiming invariance to arbitrary translations of either factor.

Proposition A.6 (Conditional Nash interpretation)**.** *Consider a compact, convex feasible utility set* $U \subset \mathbb{R}^2$ *with disagreement point* $(0, 0)$ *and at least one feasible pair with both coordinates strictly positive. On the individually rational region, the Nash bargaining solution maximizes* $u_1 u_2$ *under the standard Pareto, symmetry, positive affine covariance, and independence axioms (Nash, 1950).*

Proof. This is the two-utility Nash bargaining characterization with zero disagreement point. The general product is $(u_1 - u_1^0)(u_2 - u_2^0)$; setting $u^0 = (0, 0)$ gives the stated expression. Under positive affine transformations, the disagreement point must transform with the utilities. □

The achievable pairs $(\mathrm{Demand}(\mathbf{z}), \mathrm{Def}(\mathbf{z}; \mathcal{E}))$ for unit vectors are not shown to form such a convex feasible set. Convexifying the utility set can introduce mixtures that no individual direction realizes.

Therefore the classical theorem supplies a conditional interpretation, not a uniqueness theorem for the deployed GapScore or a proof that an additive MMR objective is invalid.

Proposition A.7 (Ambient log-concavity)**.** *Extend Demand and Defensibility using their inner-product formulas to* $\mathbb{R}^d$*. On* $\Omega = \{\mathbf{z} : \mathrm{Demand}(\mathbf{z}) > 0,\ \mathrm{Def}(\mathbf{z}; \mathcal{E}) > 0\}$*,* $\log \mathrm{GapScore}(\mathbf{z}; \mathcal{E})$ *is concave.*

Proof. Demand is affine. Defensibility is concave because it is one minus a pointwise maximum of linear functions. Their positive domains intersect in a convex set. Composing either positive concave function with the increasing concave logarithm preserves concavity (Boyd & Vandenberghe, 2004); summing the two logarithms gives the result. □

The sphere constraint is nonconvex. Ambient log-concavity therefore does not establish geodesic concavity or initialization-independent global optimization on the sphere.

A.4 Optimization: Smooth Guarantees and the Implemented Objective

At a point with a unique active neighbor, Eq. (4) is the ordinary gradient. At a tie, an active branch gives a generalized derivative choice, consistent with nonsmooth analysis (Clarke, 1983). The fixed-step implementation remains a multistart heuristic for a nonsmooth objective, in the spirit of stochastic multi-start global optimization (Rinnooy Kan & Timmer, 1987), and is related to gradient-sampling approaches for nonsmooth optimization on manifolds (Hosseini & Uschmajew, 2017); differentiability almost everywhere does not supply the smoothness assumptions needed below.

Theorem A.8 (Conditional stationarity rate for a smooth objective)**.** *Let* f *be differentiable on a compact manifold with a retraction* R*. Suppose its pullbacks obey, for the steps used, the ascent bound*

$$f(R_x(\eta \operatorname{grad} f(x))) \geq f(x) + \eta \left\|\operatorname{grad} f(x)\right\|^2 - \frac{L\eta^2}{2} \left\|\operatorname{grad} f(x)\right\|^2.$$

For $\eta = 1/L$*, some iterate among the first* T *satisfies*

$$\left\|\operatorname{grad} f(x_t)\right\|^2 \leq \frac{2L(f_{\max} - f(x_0))}{T}.$$

Thus $O(\varepsilon^{-2})$ *iterations suffice for an* ε*-stationary point under these assumptions (Boumal et al., 2019).*

Proof. Substitution of $\eta = 1/L$ gives an increase of at least $\left\|\operatorname{grad} f(x_t)\right\|^2 / (2L)$ per step. Summing over $t = 0, \ldots, T-1$ and bounding the total increase by $f_{\max} - f(x_0)$ bounds the average squared gradient. The smallest value cannot exceed that average. □

This theorem neither guarantees a global optimum nor directly covers the exact nearest-neighbor maximum in GapScore. A smooth replacement of that maximum could permit a separate analysis, but smoothing was not documented as part of the reported experiments. No convergence rate is asserted here for the original fixed-step search.

A.5 Coverage as a Separate Set Objective

Let V be a finite candidate set and let $a_{jz} = \max\{0, \langle \mathbf{u}_j, \mathbf{z} \rangle\}$. Define a normalized nonnegative coverage surrogate

$$F(S) = \sum_j f_j \max_{\mathbf{z} \in S} a_{jz}, \qquad F(\varnothing) = 0.$$

This auxiliary surrogate clips negative similarities; it is distinct from the signed diagnostic in Eq. (10).

Proposition A.9 (Submodularity and greedy coverage)**.** F *is monotone and submodular. Exact greedy maximization of its marginal gains under* $|S| \leq K$ *obtains* $F(S_K) \geq (1 - 1/e) F(S^*)$*, where* S^* *is an optimal set of at most* K *elements (Nemhauser et al., 1978). This auxiliary coverage view is related to max-sum diversification, but does not certify the implemented product objective (Borodin et al., 2012).*

Proof. For $A \subseteq B$, the current maximum for each query cannot decrease. The gain from adding z is $\max\{0, a_{jz} - \max_{y \in A} a_{jy}\}$, which cannot increase when A is replaced by B. Nonnegative weighted sums preserve this diminishing-returns property and monotonicity. By submodularity, the sum of marginal gains of elements in $S^* \setminus S_t$ is at least $F(S^*) - F(S_t)$. Since there are at most K such elements, greedy gains at least $[F(S^*) - F(S_t)]/K$. Iterating this recurrence gives $F(S_K) \geq [1 - (1 - 1/K)^K]F(S^*) \geq (1 - 1/e)F(S^*)$. □

The implemented sequential GapScore search does not maximize marginal gains of this F. Moreover, multiplying coverage by average Defensibility need not preserve monotonicity or submodularity. The theorem therefore provides no approximation factor for either the actual sequential search or SetGapScore. Repeating a direction does not increase a maximum-based coverage term, but neither does repeating identical values increase their arithmetic mean. Coverage and a mean answer different questions; their relative numerical values need not have a universal ordering.

A.6 A Conditional Connection to Source Selection

Proposition A.10 (Redundancy and a fixed marginal-relevance score)**.** *Suppose a source scorer uses* $\mathrm{MR}(s \mid E) = \lambda \,\mathrm{Rel}(s, q) - (1 - \lambda) \max_{e \in E} \mathrm{Sim}(s, e)$*, with fixed* E*, fixed relevance, and* $0 \leq \lambda < 1$*. Its score is strictly increasing in* $1 - \max_{e \in E} \mathrm{Sim}(s, e)$*.*

Proof. Writing the score as $\lambda \,\mathrm{Rel}(s, q) - (1 - \lambda) + (1 - \lambda)\,\mathrm{Def}(s; E)$ shows that its derivative with respect to Defensibility is $1 - \lambda > 0$. □

A nondecreasing inclusion probability would additionally require an inclusion rule monotone in that score, with the other candidates fixed. The GPT citation simulation is not established to satisfy these assumptions. In particular, this observation does not prove that higher GapScore must cause higher PAWC; the GEO ablation measures those quantities separately.

B Implementation Details

B.1 Shared Search and Vocabulary

Table 6: Settings used in the experimental pipeline.

Component	Setting
Encoder	`text-embedding-3-small`; 1,536 dimensions; L2 normalization
Concept decoder	Qwen2.5-3B-Instruct-Q6_K; local GGUF on Apple M2
Default filter / generator	GPT-4o-mini
Search starts	Demand direction, demand plus noise, and random direction
Search budget	15 restarts per strategy; 100 steps; step size 0.05
Answering / GEO directions	$K = 5$
Synthetic-data directions	$K = 1$ for final augmentation comparisons; $K = 3$ for set diagnostics
Vocabulary	13,409 valid words, screened to 10,000; embeddings cached in batches of 500
Concept shortlist	20 unused nearest words per candidate batch
Accepted concepts	Three per direction for answering / GEO; up to ten filtering rounds
Avoidance words	Twelve frequent words closest to the occupied-content centroid
Novelty threshold	Maximum similarity to original occupied responses below 0.7
Within-output threshold	Maximum pairwise similarity of generated sentences below 0.8
Synthetic-data target	Eight accepted examples per class per generation; at most four generation rounds

The vocabulary starts from Qwen2.5-3B-Instruct (Qwen Team, 2024) tokenizer entries restricted to alphabetic English words of length 3–15. The implementation uses the macOS dictionary at `/usr/share/dict/words` as a whitelist (234,454 dictionary entries). The 13,409 retained words are embedded once. The task-dependent screening score is

$$\mathrm{VocabScore}(w) = 0.6 \max_j \langle E(w), \mathbf{u}_j \rangle + 0.4 \max_i \langle E(w), \mathbf{e}_i \rangle \,. \tag{12}$$

The top 10,000 words are retained. The positive occupied-content term restricts the vocabulary toward the broad domain; novelty is imposed by direction search and subsequent selection. Thus the vocabulary screen and GapScore have different purposes. The coefficients are fixed heuristics in this implementation, not learned ranking weights, though the underlying idea of weighting a candidate by its similarity to occupied content echoes term-frequency-based relevance scoring in classical information retrieval (Robertson et al., 1994).

For a direction $\mathbf{z}_k$, candidates are ranked by $\langle E(w), \mathbf{z}_k \rangle$. Accepted words enter a global used-word set; rejected words can be reconsidered for other directions. The query and candidates, rather than the numerical vector, are passed to Qwen. Its decode call uses temperature zero and a 40-token output limit.

B.2 Avoidance Instructions and Output Diagnostics

When the occupied embeddings have nonzero sum, their normalized centroid is

$$\widehat{\mathbf{c}} = \frac{\sum_i \mathbf{e}_i}{\|\sum_i \mathbf{e}_i\|}.$$

The 12 frequent words nearest to this centroid form the forbidden-word prompt list. GPT-4o-mini (OpenAI, 2024) also extracts common sentence patterns from the occupied responses. The generator receives the query, a direction's concept words, and both avoidance instructions. These are prompt-level constraints rather than enforced token-level decoding constraints.

In addition to Novelty, a mean-reference statistic can be reported:

Definition 5 (CrossCollision)**.** *For a generated text g and original occupied set $\mathcal{C}$,*

$$\mathrm{CrossCollision}(g, \mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} \langle E(g), E(c_i) \rangle .$$

This measures mean similarity, not a distance. It differs from the nearest-neighbor similarity used in Novelty and from pairwise Collision within a set. Both diagnostics remain embedding-distance proxies rather than judged novelty scores: they differ from LLM-judged novelty scoring, such as blind peer-review-style feedback on creative writing (Li et al., 2026), and from bibliometric novelty measures proposed for research writing in the post-ChatGPT era (Safari & Babaei, 2026).

B.3 GEO Scoring and Synthetic-Data Filtering

The GEO scorer takes the first four competitor claims in the supplied list and appends the candidate at source position five. It does not retrieve pages from a deployed engine. Equation (9) credits all words in a sentence when it cites [5], even if it also cites another source. There is no division by the number of citations within that sentence and no subsequent normalization across source scores. These implementation details distinguish it from other position-adjusted impression measures; the reported values must be interpreted with this exact formula.

For synthetic data, concept decoding uses an expanded pool of eight words, reranked by absolute VADER compound score to retain five. This reranking is shared by random, Demand-only, Defensibility-only, and GapScore variants. The generator receives the target label separately; the final polarity check uses the sign and the thresholds described in Section 3.6. Data Accumulation does not use this same quality filter in this comparison. Equal final sample counts therefore control training-set size, but do not isolate direction choice from filtering or equalize generation cost.

Two earlier fidelity checks were tried first: GPT-4o-mini self-evaluation returned 100% fidelity for all six methods, and a seed-trained TF-IDF classifier produced unstable rankings with only ten examples per class. The adopted VADER check avoids fitting a fidelity classifier on that small seed. It is still a lexicon-based proxy and can misread contextual sentiment, especially irony or mixed polarity.

C Additional Open-Ended Generation Results

C.1 Query Groups and Topic Diagnostics

Table 7 gives the per-group breakdown of the 90 open-ended queries used in §4, grouped by question type.

Table 7: Query-group breakdown for the 90-query open-ended generation study.

Group	Queries	Baseline Collision	Mean reduction
Hypothetical	22	0.79	11.3%
Ideal	7	0.78	11.8%
Metaphor	1	0.92	17.0%
Explain	12	0.77	10.3%
Describe	20	0.78	10.1%
Imagine	10	0.77	9.9%
Philosophical	6	0.76	9.9%
Write short	12	0.76	8.9%
Reported aggregate	90	0.777	10.8%

The 90-query summary additionally reports mean final Collision 0.6915, mean GapScore 0.2012, and a mean best-direction/competitor statistic of 0.5137. It does not specify enough detail to identify the latter statistic's exact aggregation. We therefore do not rename it as Demand or use it to derive a new result. Similarly, the five-topic table reports auxiliary direction/competitor values of 0.4514, 0.4488, 0.5402, 0.3779, and 0.3638; the corresponding differences from baseline Collision are 0.2848, 0.2960, 0.2644, 0.4383, and 0.4540. These columns are not used as independently defined evaluation metrics in the main text.

Table 8: Additional five-topic diagnostics. Topic-level GapScore and the geometric search scores in Table 5 come from different tables. Vendi values describe the baseline response set; baseline Collision is given in Table 1.

Topic	Reported GapScore	Baseline Vendi	Vendi/20
Time metaphor	0.165	2.91	14.5%
Loneliness metaphor	0.196	3.16	15.8%
Redefine courage	0.221	2.58	12.9%
Memory as substance	0.220	2.36	11.8%
Hope via natural elements	0.229	2.12	10.6%
Reported mean	0.206	2.63	13.1%

The Vendi diagnostic uses 20 responses per topic and reports a mean effective diversity of 2.63 under the chosen cosine kernel. This is an effective spectral count, not an estimate of a discrete number of human-interpretable clusters. Collision values are taken from Table 1; a redundant Collision column from this Vendi diagnostic is omitted here.

C.2 Varying Content Sources and Generation Models

The mixed-set study uses three topics and four models, with five responses per model. Table 9 gives the supplementary pairwise cross-model similarities, and Table 10 gives the before/after Collision comparison for the pooled occupied set.

For the separate-source study (Table 11), each model independently supplies 20 occupied responses per topic and GPT-4o-mini remains the realization model. In contrast, the generator study fixes 20 GPT-4o-mini occupied responses and varies only the realization model. These two experiments test different dependencies and should not be pooled.

Table 9: Between-model mean similarity in the mixed-source study. Claude denotes Claude-Sonnet-4.5 (Anthropic, 2025); Gemini denotes Gemini-2.0-Flash (Google DeepMind, 2025).

Model pair	Mean similarity
GPT-4o-mini / Claude	0.660
GPT-4o-mini / Gemini	0.613
GPT-4o-mini / Qwen2.5-3B	0.576
Claude / Gemini	0.629
Claude / Qwen2.5-3B	0.578
Gemini / Qwen2.5-3B	0.554
Reported mean	0.602

Table 10: Mixed-source Collision before and after Gap Detection, for the pooled four-model occupied set.

Evaluation	Before	After	Reported reduction
Mixed occupied set, $N = 20$	0.602	0.561	7.4%

Table 11: Separate-source study, averaged over three topics. The realization model is fixed to GPT-4o-mini; novelty pass is the recorded Check-1 statistic.

Occupied-set source	Before	After	Reduction	Novelty pass	GapScore
GPT-4o-mini	0.6860	0.6098	11.2%	86.7%	0.1609
Claude-Sonnet-4.5	0.7014	0.6126	12.7%	100.0%	0.1860
Gemini-2.0-Flash	0.5709	0.5189	8.8%	100.0%	0.1791
Qwen2.5-3B	0.6028	0.5475	8.8%	100.0%	0.1627
Reported mean	0.6403	0.5722	10.4%	96.7%	0.1722

Table 12: Realization-model comparison, averaged over three topics with the occupied set fixed. GapScore differences across realization models reflect the sentences produced by each generator, not changes to the search objective.

Realization model	Collision reduction	GapScore	Novelty	Novelty pass
GPT-4o-mini	9.0%	0.1500	0.3684	80.0%
GPT-4o	11.6%	0.1579	0.4314	100.0%
Claude-Sonnet-4.5	14.8%	0.1632	0.4938	100.0%

C.3 Qualitative Examples

Table 13 reproduces the available sentence excerpts without completing the missing text. These examples show how concept words are realized; they are not human evaluations of originality. Some decoded words, such as *brave* for courage, remain close to the query semantics despite the answer filter, illustrating its imperfect selectivity. This lexical-substitution realization is narrower than style-transfer approaches that rewrite a passage's surface form while holding its content fixed (Shen et al., 2017).

Table 13: Concept words and generated excerpts for the five topics. Excerpts are truncated with ellipses for space.

Topic	Concept words	Generated excerpt
Time metaphor	poetic, nighttime, tempo	"The night hums a quiet tempo, its stars ticking softly..."
Loneliness metaphor	loner, soliloquy, joylessness	"In the quiet hours of twilight, a solitary tree stands..."
Redefine courage	fearlessness, brave, fortitude	"Grit is the lighthouse that stands tall amid the storm..."
Memory as substance	myelin, tissue, formalin	"Envision a soft, pliable tissue, reminiscent of myelin's..."
Hope via natural elements	sunlight, warmth, naissance	"A warm breeze dances through the treetops, whispering..."

D Additional GEO Results

This market-positioning framing is in the same spirit as identifying an uncontested market space (Kim & Mauborgne, 2005), though the simulation below does not verify real-world market claims. Table 14 gives the concept words and individual factors underlying the selected market results. Table 15 gives all five water-filter directions. The rows report values directly rather than recomputing each product from the rounded Demand and Defensibility figures shown.

Table 14: Selected market positions and direction factors. These are candidate concepts, not verified product specifications.

Market	Concept words	Demand	Defensibility
Water filters	campground, wilderness, hike	0.446	0.922
Protein powder	nutritional, ketogenic, peptide	0.383	1.249
Sports water bottles	insulated, lightweight, waterproof	0.509	0.782
Yoga mats	rubber, foam, handmade	0.763	0.528
Reported mean		0.525	0.870

Table 15: All five water-filter directions. GS denotes the reported direction GapScore. The highest PAWC and the highest GS $\times$ PAWC select different directions.

Direction	Concept words	Demand	Def.	GS	PAWC	Product
1	campground, wilderness, hike	0.446	0.922	0.410	0.518	0.213
2	portable, filtration, faucet	0.772	0.396	0.306	0.410	0.125
3	filter, purification, filtering	0.579	0.427	0.247	0.471	0.116
4	screened, cleansing, reservoir	0.730	0.185	0.135	0.544	0.073
5	waterproof, tank, aquatic	0.755	0.355	0.268	0.484	0.130

The per-market best-direction results and the four-method ablation in Table 3 are separate summaries. In particular, the selected-market mean GS of 0.423 should not be substituted for the ablation mean of 0.307. This simulation does not include a held-out citation check after candidate selection, source-order randomization, or repeated-run intervals. It also has only three surviving yoga-mat brands and twelve claims, limiting the coverage of that market's occupied set.

E Additional Synthetic-Data Results

E.1 Separate Multi-Generator Runs

Each generator in this study supplies both the undirected and directed texts; the embedding, Qwen decoding, GPT plausibility screen, and VADER check remain fixed. These runs use $K = 1$ and five generations. Table 16 reports the supplementary differences from Data Accumulation across the four generators, including Gemini-2.0-Flash.

Table 16: Separate four-generator synthetic-data runs. Differences are GapScore minus Data Accumulation. Negative Collision and positive Vendi differences are favorable. No accuracy-by-model values were supplied.

Generator	ΔCollision $\downarrow$	ΔVendi $\uparrow$
GPT-4o-mini	-0.019	$+1.43$
Claude-Sonnet-4.5	-0.023	$+3.77$
Gemini-2.0-Flash	-0.028	$+8.46$
Qwen2.5-3B	-0.024	-0.74

The GPT-4o-mini values in Table 16 come from this separate four-generator run rather than the main-text budget-matched comparison (-0.026 Collision, $+2.04$ Vendi), since the two experiments use different sampling procedures and should not be pooled. For Qwen, the reported accumulation Vendi is 29.78, making the deficit roughly 2.5%. Without repeated runs or intervals, it cannot be dismissed as noise or attributed to model size. The generator-level spread in this table is consistent with broader evidence that recursive synthetic-data training can degrade generative distributions (Shumailov et al., 2023) and erode diversity within a model's own outputs (Guo et al., 2023), though neither prior result was measured on this pipeline.

E.2 Direction-Set Diagnostics and Fidelity

The $K = 3$ experiment studies the directions themselves. It is distinct from the $K = 1$ end-to-end comparison. The recorded SetGapScore values and available mean-direction diagnostic values are shown in Table 17.

Table 17: Synthetic-data direction-set diagnostics ($K = 3$). A dash indicates a diagnostic not computed for that method. The two columns evaluate different aggregations, not repeated measurements of the same metric.

Method	SetGapScore	Mean direction GapScore (diagnostic)
Random direction	0.023	—
Demand only	0.199	0.199
Defensibility only	-0.029	-0.262
GapScore	0.276	0.192

GapScore's set score is approximately 44% higher than its mean-direction score, reflecting the different aggregation. This is not a 44% improvement in generated-data quality. Demand-only yields identical directions, so its coverage maximum and per-direction demand coincide. Defensibility-only also gives diagnostic values of Collision 0.259 and Vendi 32.10, showing that strong final-sample diversity can coexist with a negative direction score; these do not justify replacing downstream evaluation by SetGapScore.

A qualitative comparison of $K = 1$ and $K = 3$ for Gemini-2.0-Flash shows that spreading the generation quota across three directions improved Vendi relative to accumulation but worsened Collision and accuracy, whereas the one-direction variant improved all three. A full K-sensitivity curve is not reported here. Retained-sample matching targets eight examples per label per generation, with at most four retry rounds.

F AVAILABLE PROMPT TEMPLATES

The following templates are given as ordinary text, with braced names denoting runtime substitutions. The selection and filtering prompts are given verbatim; the final answer/synthetic-sample generation prompt is described qualitatively in §3 rather than reproduced verbatim.

F.1 QWEN CONCEPT SELECTION

Query: '`{query}`'. Candidate words: `{candidates}`.

Select up to `{top_n}` words from the candidates that would make good ANSWERS to this query—concrete, creative, unexpected words. Do NOT select words that are synonyms of the key concepts in the query. Return only the selected words, comma-separated, best first.

The documented decode settings are temperature 0.0 and `max_tokens=40`. The numerical direction is used to retrieve the candidates, not passed directly to the model.

F.2 ANSWER PLAUSIBILITY FILTER

Query: '`{query}`'. Words: `{words}`.

Which of these words could be answers to the query question? Think about what the query is asking for, then decide if synonyms of the query's key concepts should be excluded as answers. Return only those words, comma-separated. Maximum `{top_n}` words.

F.3 GEO PLAUSIBILITY FILTER

Market: `{market_desc}`. Query: '`{market_query}`'. Words: `{words}`.

Which of these words could be answers to the query question? Think about what the query is asking for, then decide if synonyms of the query's key concepts should be excluded as answers. Return only those words, comma-separated. Maximum `{top_n}` words.

F.4 CITATION SIMULATION

Write an accurate and concise answer for the given user question, using ONLY the provided summarized web search results. Every sentence must be immediately followed by an in-line citation using the format [1][2][3].

Question: `{market_query}`.

Search Results: `{source_text}`.

The source text enumerates four competitor claims followed by the candidate claim. The response is split into sentences, and sentences containing [5] contribute their position-weighted word counts to Eq. (9). The scorer uses zero-based sentence indices.

F.5 TASK-LABEL PLAUSIBILITY FILTER

Task: `{task_desc}`. Label space: `{label_names}`. Words: `{words}`.

Which of these words correspond to a valid instance of this classification task? Exclude words that are off-topic or do not correspond to any class. Return only those words, comma-separated. Maximum `{top_n}` words.

F.6 TEXT-GENERATION INSTRUCTIONS

For answering, the described prompt combines the query, three concept words, the forbidden-word list, and recurring sentence patterns, and requests one new sentence per direction. For GEO, it requests a product-claim sentence using the accepted concepts without duplicating competitor claims. For synthetic data, it requests labeled examples covering the accepted concepts, followed by the

polarity check. These descriptions establish the conditioning variables but do not substitute for exact generation prompts, decoding settings, or raw completions.