

## Article

# Assessing Early-Stage Product Innovation Opportunities from Text Co-Occurrence Networks: A Decision-Support System for the Fuzzy Front End of New Product Development

Zhiwei Wang <sup>1,†</sup> , Shengkang Gao <sup>1</sup>, Peng Lin <sup>1</sup>, Guannan Qu <sup>2,†</sup>  and Die Hu <sup>3,\*</sup>

<sup>1</sup> School of Economics and Management, Fuzhou University, No. 2 Wulongjiang North Avenue, Fuzhou University Town, Fuzhou 350108, China; wangzhiwei@fzu.edu.cn (Z.W.); 2500720062@fzu.edu.cn (S.G.); 230720090@fzu.edu.cn (P.L.)

<sup>2</sup> School of Management, Beijing Institute of Technology, 5 Zhongguancun South Street, Haidian District, Beijing 100081, China; guannanqu@bit.edu.cn

<sup>3</sup> Business School, Sichuan University, No. 29, Wangjiang Road, Wuhou District, Chengdu 610064, China

\* Correspondence: hudie@scu.edu.cn

† These authors contributed equally to this work.

## Abstract

In the fuzzy front end of innovation, firms often lack sufficient citation, market, and performance data, which limits the usefulness of outcome-based approaches to screening early-stage product innovation opportunities. To address this problem, this study develops a text co-occurrence network-based measurement system for assessing early-stage product innovation opportunities in new product development. We first preprocess idea texts through concept extraction and semantic cleaning, and then construct an integrated semantic network by combining market-related texts with ideation data. The Leiden algorithm is applied to detect latent knowledge communities in the network. Building on this structure, we assess early-stage product innovation opportunities along two complementary dimensions: cross-domain knowledge recombination, capturing the extent to which an idea draws on concept communities that are otherwise weakly connected, and network structural perturbation, capturing the degree to which an idea reconfigures existing semantic boundaries and connection patterns. Based on community entropy and modularity change, we construct a composite indicator for the ex ante assessment of early-stage ideas with stronger product innovation potential. Compared with traditional approaches relying on patent citations, market outcomes, or expert judgments, the proposed method enables earlier screening of ideas that deviate from dominant semantic trajectories and may warrant further development attention. The framework is explicitly positioned as an ex ante screening and attention-allocation tool for early-stage product innovation opportunities, not as a deterministic predictor of later market success.



Academic Editors: Tine Bertoncel and Maja Meško

Received: 30 April 2026

Revised: 9 June 2026

Accepted: 17 June 2026

Published: 1 July 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** product innovation opportunity; early-stage idea screening; text co-occurrence network; fuzzy front end; new product development; decision-support system

## 1. Introduction

The fuzzy front end (FFE) of new product development (NPD) is the phase in which firms search for opportunities, generate ideas, and define the direction of future innovation [1–5]. For firms seeking new growth and renewal, this stage is especially consequential because early screening decisions determine which ideas receive organizational attention, development resources, and follow-on investment, thereby shaping innovation portfolios

and long-term strategic trajectories [1,2]. Yet the FFE remains difficult to govern as a business process: customer needs are ambiguous, technical feasibility is uncertain, and ideas usually exist only as short textual descriptions or preliminary concept proposals rather than validated products or market-tested concepts [3–6]. Consequently, organizations often face a structural dilemma at the front end—ideas must be evaluated before reliable evidence exists, even though those early choices may strongly influence subsequent innovation outcomes.

From an innovation management perspective, the central challenge is not only generating many ideas but also assessing which early-stage ideas represent stronger product innovation opportunities. At the FFE stage, ideas are usually expressed as short concept descriptions with limited evidence regarding feasibility, market response, or commercial value [1,3,5,6]. As a result, firms often have to allocate attention and resources before reliable outcome indicators exist. Existing work also shows that early idea screening is easily confounded by novelty effects: unusual concepts may attract attention, while ideas with clearer developmental promise may be overlooked if they are weakly articulated or span established knowledge boundaries [1,4,6–8]. This challenge makes the *ex ante* assessment of product innovation opportunities a central yet underdeveloped task in front-end innovation governance.

Despite this importance, *ex ante* assessment of product innovation opportunities remains methodologically underdeveloped. Most existing empirical approaches rely on retrospective indicators such as patent citations, market trajectories, or post-launch performance [7,8]. These indicators are useful for explaining innovation outcomes *ex post*, but they become available only after an idea has matured, which makes them poorly suited to the FFE. Other approaches provide some early-stage guidance, yet they remain constrained by subjectivity, limited reproducibility, or weak scalability [1,5,6]. More broadly, although prior research suggests that high-potential innovation often emerges from boundary-spanning search and distant knowledge recombination [9–15], existing measures rarely convert this theoretical insight into an operational decision-support tool that can be embedded into front-end innovation governance. The result is a persistent gap between innovation theory and the business-process needs of early-stage idea screening.

This study addresses that gap by developing a text co-occurrence-network-based decision-support system for assessing early-stage product innovation opportunities in the FFE. Drawing on knowledge recombination theory [9–15], we argue that early-stage product or service ideas may exhibit stronger innovation opportunity signals even before market outcomes unfold. The first such signal is cross-domain knowledge recombination, which captures the extent to which an idea spans otherwise separated semantic communities. The second is network structural perturbation, which captures the extent to which an idea reconfigures the modular structure of the broader semantic network and weakens established semantic boundaries. To operationalize these mechanisms, we conduct concept extraction and semantic cleaning, construct an integrated semantic network by combining ideation texts with market-related corpus data, apply the Leiden algorithm to detect latent knowledge communities, and derive a composite product innovation opportunity score based on community entropy and modularity change. The framework is positioned as an *ex ante* screening and attention-allocation tool for early-stage product innovation opportunities, not as a deterministic predictor of later market success.

This paper makes three contributions. First, it contributes to research on front-end innovation evaluation by shifting the analytical focus from realized outcomes to the *ex ante* assessment of early-stage product innovation opportunities in the FFE. Second, it contributes methodologically by introducing a transparent and reproducible text-network framework that translates knowledge recombination and structural boundary change

into measurable semantic indicators without relying on post-launch evidence. Third, it contributes to innovation management and systems-oriented practice by offering a scalable decision-support tool that can be embedded into FFE governance for idea screening, opportunity recognition, and portfolio prioritization. Rather than treating early idea evaluation as an intuitive or purely qualitative task, the study shows how firms can use semantic-structural evidence to allocate attention and evaluation effort at the point where innovation choices are actually made.

## 2. Literature Review

To develop a front-end decision-support system for assessing early-stage product innovation opportunities, this section reviews four streams of literature that jointly motivate the proposed framework: fuzzy front-end evaluation, early-stage product innovation opportunity assessment, knowledge recombination as a mechanism of product innovation potential, and text-based innovation analytics. Rather than revisiting the general importance of the fuzzy front end, the purpose here is to clarify why existing theory and measurement approaches remain insufficient for *ex ante* screening when firms must evaluate early-stage ideas under uncertainty. We argue that the unresolved problem is not only conceptual—how to distinguish stronger product innovation opportunities from novelty, radicalness, or breakthrough status—but also methodological: how to operationalize such opportunities in a transparent, scalable, and reproducible way when the available evidence consists primarily of short idea texts rather than patents, citations, or market outcomes.

From this perspective, the fuzzy front end should be understood not only as an ideation stage but also as an innovation governance and evaluation stage in which organizations must decide which ideas deserve further attention, experimentation, and resource commitment. Existing studies consistently show that FFE decisions are made under conditions of incomplete information, weak evaluative signals, and substantial path dependence, making early screening one of the least structured yet most consequential parts of innovation management. This challenge becomes even more pronounced when the target of evaluation is early-stage product innovation opportunities, because front-end ideas rarely come with reliable evidence regarding adoption, performance, or commercialization prospects. Accordingly, the following review focuses on the theoretical and methodological foundations needed to move from retrospective explanation of realized outcomes toward *ex ante* assessment of product innovation potential in early-stage ideas.

### *2.1. Front-End Innovation Evaluation and Early-Stage Product Innovation Opportunity Assessment*

The fuzzy front end (FFE) of new product development represents the earliest stage of innovation activities, during which firms explore opportunities, generate ideas, and shape initial concept definitions [1,3]. Unlike later development stages, the FFE is characterized by ambiguous customer needs, uncertain technological feasibility, and limited availability of reliable evaluative evidence [4,5,16,17]. Consequently, decisions made in this phase are inherently subjective and strongly dependent on incomplete information [1]. Prior research consistently emphasizes that the FFE is one of the most critical yet least structured phases of innovation management, as it largely determines subsequent development trajectories and resource allocation [3,5]. From a managerial perspective, the FFE is not merely an ideation stage but also a filtering and selection stage in which organizations must decide which concepts deserve further investment [2]. Such early screening decisions often generate strong path dependence and are frequently driven by non-analytical managerial “gut feel” due to the difficulty of formalizing evaluation logic under high uncertainty [18]. This may cause some promising opportunities to be overlooked because of weak initial

performance signals, while incremental ideas may be favored because they fit existing evaluation routines [1,4]. Therefore, the central difficulty of the FFE lies not only in opportunity exploration but also in the lack of systematic and scalable mechanisms to evaluate and prioritize ideas under extreme uncertainty [2,3].

This challenge becomes even more salient when the target of screening is early-stage product innovation opportunities. In the FFE, organizations do not evaluate mature products but tentative concepts whose technical configuration, user value, and future positioning remain fluid [1,3,5]. Product innovation opportunities at this stage are therefore better understood as ideas that may open new product directions, use scenarios, or value propositions, rather than as already validated outcomes. Under these conditions, firms must decide which ideas deserve further experimentation and resource commitment, even though available evidence is partial, noisy, and text-based [2,6,18].

A critical implication is that early-stage product innovation opportunities cannot be assessed solely through retrospective indicators such as diffusion patterns, market substitution, or long-term performance [7,8]. To make opportunity assessment feasible in the FFE, it is necessary to shift the analytical focus from realized outcomes to semantic and structural signals embedded in idea texts. In this study, product innovation opportunities are defined as front-end product or service ideas that show stronger potential for new product development because they recombine heterogeneous knowledge elements, connect otherwise separated semantic communities, or introduce concept structures that merit further experimentation. Recent work on weak signal detection and emerging topic identification has similarly emphasized the value of extracting early indicators from textual and bibliometric data [19,20]. However, scalable measurement tools for screening large volumes of early-stage ideas remain underdeveloped, which leaves a major gap between theoretical interest and managerial applicability [1,5,6,21].

## 2.2. Knowledge Recombination as a Mechanism of Product Innovation Potential

Knowledge recombination theory provides a fundamental explanation for how innovation emerges through the restructuring of existing knowledge elements. Its intellectual roots can be traced to Schumpeter's notion of "new combinations," which argues that innovation does not primarily originate from the creation of entirely new elements but from novel combinations of existing components [9,10]. This perspective suggests that the core mechanism of innovation lies in how knowledge components are selected, connected, and reorganized into new configurations [9]. Within this framework, innovation outcomes depend not only on the availability of diverse knowledge sources but also on the integration patterns among these sources. Prior research demonstrates that cross-domain or distant knowledge recombination expands the combinatorial search space and increases the likelihood of producing novel and high-impact outcomes [12–15]. Boundary-spanning search enables innovators to escape local search traps and access heterogeneous components that may support breakthrough innovation [12]. Similarly, studies on search behavior highlight that exploration across domains is more likely to produce radical innovation than deep exploitation within a single domain [14]. Empirical evidence from scientific and technological domains also indicates that atypical combinations of knowledge elements are disproportionately associated with high-impact results, suggesting that novelty emerges from unusual integration patterns rather than from isolated novelty of individual components [15].

However, while the diversity of knowledge sources provides a potential space for innovation, its effect is not consistently stable. As the number of knowledge components increases, the heterogeneity and cognitive distance among these elements also widen, which may broaden the combinatorial space while simultaneously increasing the difficulty

of knowledge comprehension and integration [11]. Related research has highlighted a clear complexity trade-off in the recombination process: diversified knowledge inputs provide more potential recombination pathways, yet excessive heterogeneity may reduce the likelihood of successful integration and weaken innovation performance [13,14]. Importantly, recombination is not simply about diversity but about structural reconfiguration. New combinations may destabilize existing knowledge structures and require non-trivial integration [11,12]. From this view, product innovation potential can be understood as the ability of an idea to bridge distant domains and create new relational structures that challenge established boundaries [11,15]. Therefore, promising product innovation opportunities are likely to exhibit two essential characteristics: (1) cross-domain integration, reflecting recombination across heterogeneous knowledge communities [11,12], and (2) structural boundary change, reflecting the potential to reconfigure existing knowledge networks [15]. These two dimensions jointly capture both the compositional and structural aspects of recombination and together provide a theoretically grounded basis for assessing product innovation potential at an early stage. However, to apply this mechanism in the FFE, a methodological approach is required that can capture recombination patterns and structural change from the limited evidence available in early-stage idea descriptions [3,5]. Text-based innovation analytics offers a feasible pathway for such operationalization [22].

### *2.3. Text-Based Innovation Analytics and Existing Measurement Approaches*

In the FFE stage, innovation ideas are often recorded primarily as short concept descriptions, informal documentation, or unstructured textual proposals, rather than mature prototypes or market-tested products [1,3,6]. Compared with later-stage data such as patents or market reports, textual idea descriptions provide the earliest available evidence of how innovators frame problems, articulate solutions, and integrate knowledge components [5]. Therefore, text analytics has increasingly been adopted to extract latent innovation signals and enable early-stage evaluation [22]. Among various text-based approaches, co-occurrence networks provide a particularly suitable representation for innovation analysis. A co-occurrence network is constructed by linking terms that appear together within a defined textual context, thereby capturing semantic associations and latent conceptual structures [23]. In such networks, nodes represent concepts and edges represent co-occurrence relationships, which collectively form a semantic landscape of knowledge elements embedded in a corpus. Compared with purely frequency-based keyword analysis, network representations preserve relational information and thus enable structural interpretation of innovation content. Subsequent bibliometric studies further demonstrated that co-occurrence networks can effectively represent the intellectual structure of a domain and uncover thematic clusters [24,25]. Furthermore, community detection algorithms provide a powerful mechanism for identifying clustered semantic domains in co-occurrence networks [26,27]. These advances indicate that text-based network approaches provide a promising pathway for identifying early innovation signals in contexts where conventional performance indicators are unavailable [19,20,22].

At the same time, research on innovation measurement has gradually evolved from early subjective evaluation approaches toward more objective analytical methods grounded in citation relationships, semantic content, and structural characteristics of knowledge systems. Early approaches relied mainly on expert assessment and subjective scoring frameworks. With the increasing availability of large-scale innovation data, citation-based indicators offered stronger objectivity and clearer operationalization [7,8], but their major limitation lies in temporal lag, since citation structures emerge only after diffusion. More recently, text-based approaches have been introduced to measure technological novelty and semantic distance directly from innovation content [7,28]. These approaches offer

a promising direction by extracting signals directly from innovation content rather than relying solely on diffusion-based evidence. However, a key challenge is that semantic novelty does not necessarily imply valuable product innovation opportunities, and text-based novelty measures often struggle to distinguish meaningful opportunity signals from ideas that are merely unusual in wording or concept enumeration [7,28]. In parallel, structural approaches emphasize that innovation should also be assessed by its impact on the overall configuration of knowledge systems [29–31]. Compared with methods focusing on local relationships, structural approaches enable a more holistic characterization of how ideas reshape knowledge systems, although their performance remains sensitive to text preprocessing and network construction procedures [31]. Overall, existing research has established multiple measurement pathways for early innovation evaluation, including subjective evaluation frameworks, citation indicators, text-based semantic measures, and structural network-based approaches [7,8,28–31]. Yet most measurement approaches are designed for patents or scientific publications and depend on citation or structured textual evidence embedded within established knowledge systems [32,33]. This limits their applicability in the fuzzy front end, where innovation inputs exist primarily as early-stage idea descriptions without diffusion traces [3,5].

#### *2.4. Research Gap and Analytical Framework*

The above review highlights that early innovation measurement has progressed from subjective expert-based evaluations toward more objective approaches based on citation structures, semantic content, and knowledge network configurations [7,8,28–31]. However, despite these methodological improvements, a key limitation remains: most existing indicators are outcome-oriented and rely on diffusion traces embedded in established knowledge systems, such as citation pathways, patent metadata, or mature textual records [3,5]. As a result, these methods are difficult to apply to the fuzzy front end, where innovation concepts exist mainly as early-stage idea descriptions and lack observable market impact or citation-based evidence [3,5]. Moreover, the literature reveals that different measurement streams capture only partial aspects of innovation potential. Citation-based indicators primarily focus on pathway displacement or novelty, text-based semantic approaches are more sensitive to content differentiation, and structural network approaches examine reconfiguration of knowledge communities [7,8,28–31]. Consequently, current measurement approaches lack an integrated framework that simultaneously captures both the compositional diversity of knowledge recombination and the structural reconfiguration effects that indicate broader product innovation potential while also remaining cautious about the distinction between early-stage opportunities and realized outcomes.

To address these limitations, this study proposes a two-dimensional measurement system grounded in knowledge recombination theory and operationalized through text co-occurrence networks. Specifically, we argue that early-stage product innovation opportunities in the fuzzy front end can be characterized by (1) cross-domain knowledge recombination, which reflects the extent to which an idea integrates heterogeneous knowledge communities, and (2) network structural perturbation, which reflects the extent to which an idea induces reconfiguration of semantic boundaries and relational structures. By integrating these two dimensions into a unified measurement framework, this study provides an *ex ante* and scalable screening and attention-allocation tool for evaluating early-stage product innovation opportunities using only early-stage textual evidence. This contributes to innovation management research by bridging FFE screening practice and knowledge-recombination-based opportunity evaluation and by offering a transparent and reproducible analytical tool to support opportunity assessment under uncertainty.

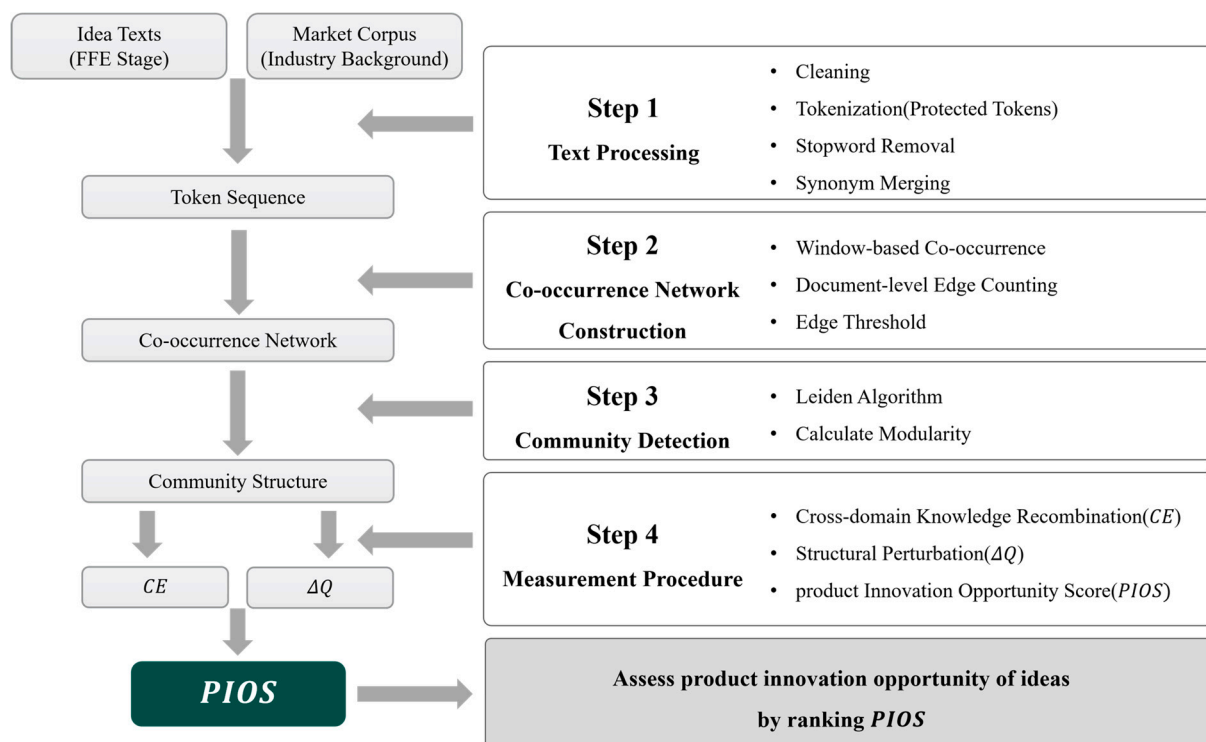
### 3. Methodology

#### 3.1. Framework Overview

Building on the literature and research gap identified in Section 2, this study argues that early-stage product innovation opportunities in the FFE can be assessed through a process-oriented evaluative lens grounded in semantic structure and knowledge recombination. Because conventional outcome indicators are not yet available in the FFE, the most accessible evidence for systematic evaluation lies in the semantic structure of idea texts. We therefore develop a text co-occurrence-network-based framework that transforms unstructured ideas into analyzable semantic relations and uses those relations to assess product innovation opportunity potential in the FFE of NPD. The framework is positioned as an ex ante screening and attention-allocation tool for early-stage product innovation opportunities, not as a deterministic predictor of later market success. Its purpose is to highlight ideas whose semantic-structural patterns may warrant further attention because they suggest stronger potential for product development, portfolio experimentation, or concept refinement.

The purpose of the method is to analyze how ideas are positioned within a broader semantic knowledge system and to assess which ideas exhibit stronger product innovation signals at an early stage. Compared with traditional front-end screening approaches that rely heavily on manual evaluation, the proposed framework provides a structured and quantifiable basis for idea assessment. In this sense, it functions not only as a measurement model but also as a lightweight decision-support tool for innovation governance in the FFE.

This method consists of four modules: text processing, network construction, community detection, and measurement procedure. Together, these modules convert raw idea texts into a semantic network and then quantify both the compositional breadth of each idea within the knowledge space and its potential to reconfigure existing semantic structures. Figure 1 presents the overall analytical flow.



**Figure 1.** Flowchart of identifying product innovation opportunity assessment method.

### 3.2. Step-1: Text Processing

In the FFE, ideas are usually recorded as unstructured text [6]. This format is easy to collect, but it often contains vague descriptions, informal expressions, and mixed terminology [34]. If left untreated, such features can introduce noise into subsequent natural language processing and distort the network structure used for measurement.

Therefore, it is necessary to conduct text cleaning and preprocessing. For ease of exposition, the following notations are introduced. Let  $I = \{1, 2, \dots, N\}$  denote the set of collected idea texts,  $N = |I|$  represent the total number of idea texts, and  $i \in I$  denote each individual idea text.

First, the collected idea texts must be cleaned. The original texts may contain colloquial expressions, spelling errors, and redundant punctuation marks [35]. In this step, invalid content such as meaningless words and special symbols should be removed. Spelling errors should also be corrected to ensure that the idea text information is more standardized and usable. Meanwhile, it should be noted that the semantic meaning of the text should not be affected during this process.

Tokenization is then used to transform each idea text into discrete units for later analysis. In our context, tokens serve as the basic semantic units from which concept co-occurrence relations are constructed.

Considering the unstructured and non-standard nature of idea texts, rule-based tokenization is necessary. This helps ensure that the core information of idea texts is not distorted during tokenization. In idea texts, many idea concepts appear in the form of multi-word expressions (e.g., “artificial intelligence”, “digital twin”, “carbon footprint”). If conventional tokenization is directly applied, it may lead to over-segmentation and semantic loss. To ensure the semantic integrity of tokens and interpretability of research, it is necessary to identify and preserve these multi-word terms as protected terms [36]. This can be achieved by combining noun phrase chunking with statistical filtering.

FFE idea texts are not always standardized. They frequently include redundant expressions and overly generic terms (e.g., “product”, “user”, and “service”) that contribute little analytical value and may distort the semantic network. A stopwords mechanism is therefore used to filter out such low-information tokens [37].

In addition, FFE idea texts are also inherently ambiguous. Emerging concepts often lack standardized terminology, and the same concept may appear in multiple synonymous or abbreviated forms. To improve network stability, synonym merging is introduced to map alternative expressions onto unified concept labels [38].

After the above preprocessing steps, the original unstructured idea texts are transformed into structured sequences of core concept tokens. These sequences concisely and accurately represent the innovation concepts expressed in the original texts. They will serve as the basis for the subsequent construction of the semantic co-occurrence network.

Formally, we denote the token sequence of each processed idea text as  $T_i = \{t_{ij}\}$ , where  $t_{ij}$  represents the  $j$ -th token derived from the  $i$ -th text, and let  $T = \{T_i\}$  represent the set of all processed idea texts.

### 3.3. Step-2: Network Construction

After text processing, we construct a semantic network to represent the knowledge structure embedded in the idea corpus. In this network, nodes denote concept tokens and weighted edges denote co-occurrence relations [23], thereby preserving the combinatorial relationships among innovation concepts in the FFE.

Many recent studies represent text data through semantic similarity networks based on embedding vectors [7,28,39]. Although useful for analogy and inference, such approaches may be less suitable in small, domain-specific corpora because they depend heavily on

pretrained representations and often blur industry-specific or emerging concepts [40]. For the present task, our goal is not general semantic similarity but the explicit representation of concept combinations within early ideas.

For this reason, we adopt a token co-occurrence network. This approach directly captures the local co-appearance of concepts in idea texts, requires less data, remains highly interpretable, and is well-suited to domain-specific ideation settings in which combinatorial relations are more informative than generic semantic proximity.

According to previous studies on co-occurrence networks, a weighted undirected co-occurrence network can be constructed based on the dataset  $T$ :

$$G = (V, E, W) \quad (1)$$

where  $V$  represents the node set of the network. It consists of unique tokens  $v \in V$ , meaning that there are no duplicated token vertices.  $E$  represents the undirected edge set,  $e_{ij} \in E$  indicating the co-occurrence relationship between token vertices  $v_i$  and  $v_j$ .  $W$  represents the edge weight matrix.  $w_{ij} \in W$  indicates the corresponding weight of undirected edge  $e_{ij}$ . It can be used to represent co-occurrence strength between tokens. If there is no edge between two tokens, it means that their corresponding weight is zero.

Next, co-occurrence strength needs to be defined. In co-occurrence network construction, co-occurrence is commonly defined in two ways: document-level co-occurrence and window-based co-occurrence. Document-level co-occurrence means that if two tokens appear in the same idea text, they are considered to have a co-occurrence relationship [41]. Window-based co-occurrence requires that two tokens appear within a fixed sliding window [23].

Compared with document-level co-occurrence, window-based co-occurrence imposes stricter positional requirements. By constraining the window size, co-occurring tokens must be close to each other in the text. Because the processed idea texts retain a moderate number of tokens and because local context is important for capturing idea-level semantic relations, this study adopts window-based co-occurrence. The window size is denoted as  $S$ , and the window sliding step is set to 1 by default to make the window move more smoothly.

Window-based co-occurrence can nevertheless overcount repeated token pairs within a single document. Because our interest lies in whether a concept pair is expressed in an idea rather than how often it is repeated rhetorically, we use document-level deduplication when counting within-document pairs. This reduces mechanical inflation caused by text length or listing behavior [42]. That is, in the same idea text, regardless of how many times the same token pair appears, it is counted only once.

A binary variable is used to indicate whether the token pair  $(v_j, v_k)$  has a co-occurrence relationship in the idea text  $i$ :

$$R_{jk}^{(i)} = \begin{cases} 1, & \text{if } (v_j, v_k) \text{ occur within at least one window in text } i \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Finally, the edge weight is defined as the co-occurrence frequency of the token pair across dataset  $I$ :

$$w_{jk} = \sum_{i=1}^N R_{jk}^{(i)}, \quad i \in I, \quad e_{jk} \in E \quad (3)$$

Furthermore, to meet the structural requirements of subsequent computations, this model adopts an edge filtering strategy. Specifically, by applying an edge weight threshold

$\tau$ , a more robust backbone network with fewer spurious correlations can be extracted [43]. Only edges satisfying the following condition are retained:

$$w_{jk} \geq \tau \quad (4)$$

If the number of occurrences of a token pair in the entire network is less than this threshold, its edge weight is set to zero. This indicates that there is no co-occurrence relationship between the two tokens. This requirement helps maintain a stable semantic backbone of the network and improves the reliability of subsequent community detection and structural measurement. Then, the semantic backbone network is defined as  $G_B$ .

The values of the window size  $S$  and the edge weight threshold  $\tau$  can be determined according to the characteristics of the collected texts.

Through the above procedure, the processed discrete token set  $T$  is reorganized into a co-occurrence network  $G$  with explicit co-occurrence relationships. The associations between tokens are extracted as the basis for subsequent network analysis. Further quantitative evaluation of idea texts can be conducted through the co-occurrence relations in the network.

### 3.4. Step-3: Community Detection

Constructing the co-occurrence network provides only an initial representation of semantic structure. To reveal the latent organization of the knowledge space, we next apply community detection, which enables deeper structural analysis of the semantic network.

Under the knowledge recombination theory, innovation is regarded as a process of new combinations of existing knowledge components [9]. High-impact innovation often originates from distant search and unconventional combinations across domains [8,11]. These unconventional combinations usually span knowledge domains with large cognitive distance [15]. They are often associated with long-path search and non-adjacent knowledge recombination [44]. Therefore, identifying the boundaries and structures of knowledge domains is a prerequisite for evaluating whether idea texts involve cross-domain combinations.

Community detection identifies groups of concepts that are densely connected internally but sparsely connected externally. In the present study, these communities are interpreted as latent knowledge domains or thematic clusters within the idea texts [42].

Through community detection, the co-occurrence network can divide different tokens into different co-occurrence communities based on co-occurrence relations. Specifically, the node set ( $V$ ) of the weighted undirected co-occurrence network ( $G$ ) obtained in Step-2 is partitioned into  $M$  communities. The community set is denoted as  $C$ :

$$C = \{C_1, C_2, \dots, C_M\} \quad (5)$$

Each node is assigned to a unique community, and communities are mutually exclusive:

$$V = \bigcup_{m=1}^M C_m \text{ and } C_a \cap C_b = \emptyset, \forall C_a, C_b \in C \quad (6)$$

Based on the properties of communities, tokens in the same community are more likely to co-occur. Their combinations are therefore more conventional. In contrast, combinations between tokens from different communities indicate associations that are relatively uncommon in the existing knowledge system. They are also structurally distant. Such combinations can be considered relatively unconventional recombinations. Under the knowledge recombination theory, they are more likely to have product innovation opportunities [44].

In community detection, modularity is commonly used to evaluate network community partitions [27]. The core idea is that in a good community partition, edges within communities should be significantly higher than the expected value under a random null model. Therefore, modularity becomes the objective function of many community detection algorithms.

The Leiden algorithm stands out in community detection due to its refinement step and improved modularity optimization. Compared with other algorithms (e.g., Louvain algorithm), the Leiden algorithm not only runs faster but also provides stronger connectivity and stability [26]. It reduces the risk of fragmented or disconnected communities. The modularity function of the Leiden algorithm is defined as:

$$Q = \frac{1}{2W_c} \cdot \sum_{j,k} \left( w_{jk} - \gamma \cdot \frac{s_j s_k}{2W_c} \right) \cdot \delta(c_j, c_k) \quad (7)$$

where:

- $w_{jk}$  represents the weight of the edge between node  $v_j$  and node  $v_k$ ;
- $s_j$  represent the weighted degree of node  $v_j$ :  $s_i = \sum_j w_{ij}$ ;
- $W_c$  represents the sum of all edge weights in the network:  $W_c = \frac{\sum_{i,j} w_{ij}}{2}$ ;
- $c_j$  represents the community label of node  $v_j$ :  $c_j = m$ , when  $v_j \in C_m$ ;
- $\delta(c_j, c_k)$  is an indicator function, which equals 1 if node  $v_j$  and node  $v_k$  belong to the same community, and 0 otherwise;
- $\gamma > 0$  is a resolution parameter used to adjust the connection density within and between communities. A larger  $\gamma$  leads to more communities that are smaller in size and more densely connected.

Since community partition serves as the foundation for subsequent quantitative evaluation, the robustness of Leiden community detection is an important reason for selecting this algorithm.

Through community detection, the co-occurrence network is divided into multiple interpretable knowledge communities. This supports objective analysis of the co-occurrence network from the perspective of community distribution and community changes.

### 3.5. Step-4: Measurement Procedure

In the FFE, relevant outcome evidence has not yet emerged. Ex post indicators such as patent citations, diffusion dynamics, or market performance are therefore unavailable for assessing product innovation opportunities. Under these conditions, the explicit conceptual knowledge embedded in idea texts becomes one of the few observable and analyzable sources for early-stage evaluation.

From a systems perspective, each idea can be treated as an element within a broader knowledge structure. Its product innovation potential depends not only on its internal composition but also on its relational position and its capacity to alter existing semantic structures. We therefore measure product innovation opportunity from both intrinsic and relational viewpoints.

Specifically, this study operationalizes product innovation opportunity measurement along two complementary dimensions: cross-domain knowledge recombination and structural perturbation.

#### 3.5.1. Cross-Domain Knowledge Recombination

Cross-domain knowledge combination primarily captures characteristics of unconventional recombination in creative ideas at the level of knowledge elements. From the perspective of knowledge recombination theory, innovation is regarded as a process of

recombining existing knowledge components, and high-impact innovations often arise from distant search across domains and unconventional combinations. Compared with incremental improvements within a single domain, combinations that span multiple knowledge boundaries are more likely to deviate from mainstream trajectories, thereby fostering potentially product innovation opportunities [15].

Therefore, during the FFE stage, if an idea simultaneously draws upon multiple relatively separated knowledge communities, it is more likely to exhibit strong cross-domain recombination characteristics and product innovation potential at the semantic level.

Accordingly, this model employs a community entropy metric to measure the distributional dispersion of creative concepts across different communities, serving as a quantitative characterization of cross-domain knowledge combination.

For an idea text  $i$ , let  $n_m^{(i)}$  be the number of tokens in community  $m$  within its token sequence  $T_i$ , and let  $N_i = \sum_{m=1}^M n_m^{(i)}$  denote the total token count. The proportion of tokens in community  $m$  is then defined as:

$$p_{im} = \frac{n_m^{(i)}}{\sum_{m=1}^M n_m^{(i)}} \quad (8)$$

Then, cross-domain knowledge combination is defined as:

$$CE_i = - \sum_{m=1}^M p_{im} \log_2 p_{im} \quad (9)$$

When tokens of an idea are predominantly concentrated within a single community,  $CE$  approaches 0, indicating that the idea primarily involves incremental improvements within an existing domain. Conversely, when tokens of an idea are distributed across multiple distinct communities in a dispersed manner,  $CE$  will increase, suggesting that the idea draws upon a broader range of knowledge sources and exhibits stronger cross-domain combinatorial characteristics.

Unlike simply counting the number of categories, this metric more effectively captures cross-domain recombination characteristics in the knowledge space. It also provides a structural basis for assessing product innovation opportunities. A higher  $CE$  indicates that the creative idea spans more knowledge boundaries at the semantic level. The knowledge sources are more dispersed, making it more likely to produce distant knowledge recombination and deviate from established technological trajectories.

By adopting the community entropy method, the cross-community distributional scope of an idea can be objectively identified from data. This is achieved through the community structure of the overall semantic network, rather than relying on subjective judgments by researchers.

### 3.5.2. Structural Perturbation

Cross-domain recombination alone is not sufficient to identify higher-potential product innovation opportunities. Some ideas may mention multiple domains but merely concatenate concepts without meaningfully changing the structure of the knowledge space.

Prior research suggests that high-impact or breakthrough themes are often associated with visible changes in knowledge network structure, which can be captured through structural perturbation [8,31]. We therefore introduce a network structural perturbation metric to measure the extent to which an idea reconfigures existing semantic boundaries.

Unlike cross-domain combination, structural perturbation emphasizes whether an idea alters the modular structure of the overall semantic network through establishing

cross-community connections. New knowledge combinations can produce restructuring effects on the overall network structure by changing inter-community connection patterns. This confers stronger product innovation potential [8].

Modularity is an important metric for measuring community structure partitions. It can also be used as an outcome variable to observe changes in network structure. That is, when an idea with stronger product innovation potential is added into the community co-occurrence network, it may not strengthen the community structure. Instead, it may weaken community boundaries and lead to a decrease in modularity. However, if an idea does not have strong product innovation potential, adding it to the network may strengthen internal community connections. This may make the community partition clearer and lead to an increase in modularity, or it may have no effect on community structure.

Therefore, modularity change can be used to infer whether an idea contains stronger product innovation opportunities signals. It serves as one of the criteria for opportunity assessment in the FFE stage.

This model adopts an idea removal strategy to compare modularity changes and calculate the structural perturbation degree of an idea. The optimal modularity ( $Q_0$ ) obtained in Step-3 using the Leiden algorithm is used as the baseline value. Then, for an idea text  $i$ , the edge weights generated by this idea are removed from the co-occurrence network to obtain a new network  $G_{-i}$ . Under the original community partition, the modularity of the new network ( $Q_{-i}$ ) needs to be recalculated. The structural perturbation can be defined as the modularity change:

$$\Delta Q_i = Q_{-i} - Q_0 \quad (10)$$

When  $\Delta Q_i$  is significantly greater than 0, it indicates that the modularity increases significantly after removing the idea  $i$ . This suggests that the idea weakens community boundaries. It also implies that the idea has stronger perturbation and reconstruction potential on the overall knowledge structure.

In contrast, if  $\Delta Q_i$  approaches 0 or even becomes negative, it suggests that the idea mainly occurs within communities. In such cases, the idea has limited influence on the overall structure, or it may even make the overall structure more stable.

Compared with measurement methods based on the number of concepts or the distribution of low-frequency words, this metric focuses on capturing the reconstruction effect of an idea on the existing knowledge connection structure. Thus, it can more accurately reflect the structural influence of the idea in the knowledge network.

### 3.5.3. Product Innovation Opportunity Score

The above two metrics capture complementary aspects of product innovation opportunities in the FFE stage. Specifically,  $CE$  reflects whether an idea integrates concepts from multiple knowledge communities, while  $\Delta Q$  reflects the extent to which an idea reshapes the modular structure of the semantic network. Therefore, this model integrates the two metrics to construct a composite product innovation opportunity score ( $PIOS$ ) for each idea.

Although document-level deduplication was applied during network construction to avoid repeated counting of identical concepts within a single idea, ideas containing numerous distinct concepts may still generate inflated scores due to larger semantic participation. To alleviate potential biases caused by concept enumeration and the structural participation scale, different adjustment strategies were applied to the two indicators according to their underlying mechanisms.

For the  $CE$  indicator, a logarithmic length adjustment was introduced:

$$CE_{i,adj} = \frac{CE_i}{\log(N_i + 1)} \quad (11)$$

where  $N_i$  denotes the total token count of the idea text  $i$ . This adjustment was adopted because entropy-based measures naturally increase as more tokens participate across semantic communities. The logarithmic transformation provides a moderate penalty for excessively long or enumeration-style ideas while preserving the ability of  $CE$  to capture cross-domain semantic diversity.

Unlike  $CE$ , the  $\Delta Q$  indicator reflects structural perturbation in the semantic co-occurrence network and is primarily influenced by the scale of semantic relations contributed by an idea. Therefore,  $\Delta Q$  was normalized by the total removed edge weight associated with the focal idea:

$$\Delta Q_{i,norm} = \frac{\Delta Q_i}{R_i} \quad (12)$$

where the structural participation strength of idea text  $i$  is defined as:

$$R_i = \sum_{(v_i, v_j) \in E_i \cap E_G} w_{ij} \quad (13)$$

where  $E_i$  denotes the semantic edge set generated by idea text  $i$ ,  $E_G$  denotes the edge set of the backbone semantic network. Because document-level deduplication was adopted during network construction, each semantic relation within a single idea contributes at most once to  $E_i$ . In addition, only semantic relations retained in the backbone network  $G_B$  contribute to  $R_i$ . Therefore,  $R_i$  reflects the overall structural participation strength of idea text  $i$  in the semantic network rather than repeated concept occurrences.

This normalization evaluates the structural perturbation efficiency per unit of structural participation, thereby reducing the influence of lengthy or concept-listing ideas that contain many semantic concepts but contribute limited structural reconfiguration.

Since the adjusted  $CE_{i,adj}$  and normalized  $\Delta Q_{i,norm}$  metrics are measured on different numerical scales, they need to be standardized prior to aggregation. This model applies Z-score standardization to transform each metric into a comparable scale-free value. For a metric  $X_i$  of idea  $i$ , the standardized score is computed as:

$$Z_i = \frac{X_i - \mu_X}{\sigma_X} \quad (14)$$

where  $\mu_X$  and  $\sigma_X$  denote the mean and standard deviation of metric  $X$  across all ideas, respectively.

Z-score standardization is adopted because it provides a normalized representation with zero mean and unit variance, allowing the composite score to reflect the relative deviation of each idea from the overall distribution. This method improves interpretability and enables robust comparison of product innovation opportunity potential across ideas.

Accordingly, the standardized cross-domain recombination score and the standardized structural perturbation score are denoted as  $CE'_i$  and  $\Delta Q'_i$ . The  $PIOS$  is then defined as:

$$PIOS_i = \alpha \cdot CE'_i + (1 - \alpha) \cdot \Delta Q'_i \quad (15)$$

where  $\alpha \in [0, 1]$  is a weighting parameter that determines the relative contribution of the two dimensions. A higher  $\alpha$  indicates greater emphasis on cross-domain knowledge recombination, whereas a lower  $\alpha$  indicates greater emphasis on structural perturbation.

Given that there is currently no established prior theory or empirical evidence supporting differential weighting between the two dimensions, assigning unequal weights may introduce subjective bias. Therefore, following the principle of Occam's razor and the standard practice in composite indicator construction [45–47], this study adopts an equal-weight linear aggregation strategy to ensure transparency and robustness. Accordingly,  $\alpha$  is set to 0.5, and the *PIOS* is computed as:

$$PIOS_i = 0.5 \cdot CE'_i + 0.5 \cdot \Delta Q'_i \quad (16)$$

Finally, ideas are ranked according to their *PIOS*<sub>*i*</sub>. Ideas with higher *PIOS* are considered to exhibit stronger product innovation opportunity potential in the FFE stage.

In summary, the proposed model provides a transparent and reproducible measurement method for product innovation opportunities, enabling objective evaluation of innovation ideas in the FFE stage. By ranking ideas based on their *PIOS*, the method can identify those most deserving of further evaluation and experimentation.

## 4. Results

To validate the feasibility and interpretability of the proposed measurement framework, this section conducts an illustrative case study in the domain of tea products and tea-related services by following the procedures described in Section 3. Each text was first cleaned and transformed into a set of tokens. A sliding-window co-occurrence strategy was then applied to capture local semantic associations in each idea. Then, all co-occurrence relations were aggregated to construct a global semantic co-occurrence network. Finally, *PIOS* is calculated for each idea based on this network and product innovation opportunities are assessed accordingly. The tea domain was selected because it simultaneously contains traditional consumption patterns and emerging innovation trends, providing an appropriate context for observing diverse semantic recombination in the fuzzy front-end stage.

### 4.1. Data Description and Preprocessing

The idea dataset in this case was collected from an on-site team-based experiment organized by our research team. Participants included MBA student teams and undergraduate student teams. All participants had received innovation-related training and were able to complete structured ideation tasks. Before the experiment, participants were randomly assigned to teams of three to four members. Each team was required to collaboratively generate ideas around the topic of “tea products and tea-related services” and submit a complete textual description as the final output.

To ensure data quality, several invalid samples were removed based on abnormal records observed during the experiment. After data cleaning, 187 valid teams were retained. The final dataset provides a moderately sized and high-quality idea corpus, which is suitable for validating the proposed text co-occurrence network-based measurement framework.

In addition to the experimental idea dataset, this case further incorporated a tea industry market corpus collected from publicly available secondary sources in March 2026 in order to enhance the external reference of the semantic network structure and improve the representativeness of the industry semantic background. The market corpus consisted of 300 curated tea-industry-related semantic records covering representative concepts associated with tea beverages, health-oriented tea products, tea consumption scenarios, tea-related devices, lifestyle services, and emerging cross-domain tea applications.

On the one hand, the market corpus provides a relatively mature and stable conceptual association structure for the tea industry. This ensures that the semantic network is not solely determined by the experimental ideas, thereby reducing structural bias caused by

limited sample size or insufficient semantic coverage. On the other hand, the market corpus serves as an external semantic baseline, helping the detected community structure better approximate the real semantic ecology of the industry. This provides a more robust structural reference for subsequent measurements of cross-domain recombination and structural perturbation.

It should be emphasized that all indicators proposed in this study are calculated only based on the experimental idea texts. The market corpus is used solely as supplementary contextual information during the semantic network construction stage rather than as an evaluation object. Accordingly, the preprocessing and network integration procedures applied to the market corpus were kept consistent with those used for the experimental idea texts.

Since FFE idea texts are typically unstructured, they often contain redundant descriptions, template-based expressions, frequent generic terms, and mixed use of synonymous expressions [6,35,38]. If a network is constructed directly from raw texts, semantic noise may be amplified and may further interfere with indicator calculation. Therefore, following the preprocessing procedure proposed in Section 3, both the experimental idea corpus and the market corpus were processed using a unified preprocessing and concept extraction procedure. This ensures that network nodes can effectively represent key semantic components of the idea structure.

Specifically, since all texts in this dataset are Chinese, we applied the Chinese tokenization tool jieba (version 0.42.1), which supports domain dictionary expansion and flexible segmentation for Chinese texts [48]. During tokenization, part-of-speech tagging was used to retain major nouns and action-oriented verbs with operational meaning (e.g., “blending”, “cold brewing”, and “cold extraction”), so as to extract core concepts that are more likely to reflect the structural characteristics of the idea network.

To improve tokenization accuracy, a tea-industry-specific protected term dictionary and a list of other protected expressions were constructed. These resources were used to preserve multi-word expressions that frequently appear in tea-related innovation concepts (e.g., “ready-to-drink tea”, “fruit tea”, “gift box”, “points redemption”), including tea product formats, consumption scenarios, and technology-related phrases. This step is crucial because incorrect tokenization may split meaningful concepts into isolated tokens, which may distort the resulting network structure.

In addition, a stopword list was applied to remove high-frequency but low-information terms (e.g., “product”, “service”, and “user”), which commonly appear in generic innovation templates. Moreover, synonym merging was conducted to reduce semantic redundancy caused by alternative expressions and abbreviations, thereby improving the statistical stability of nodes in the co-occurrence network.

The descriptive statistics of the preprocessing results are reported in Table 1. In the raw tokenization stage, each idea text contained an average of 217.63 tokens (Min = 37, Max = 870), indicating substantial variation in text length across teams. After stopword filtering, the average number of tokens decreased to 139.76, suggesting that a large number of generic low-information terms were effectively removed. In the final concept extraction stage, each text was further compressed into a core concept set, retaining an average of 20.78 concepts. Compared with the raw tokenization results, the final retained concept size accounts for approximately 9.6% of the original tokens. This demonstrates that the proposed preprocessing mechanism can significantly reduce redundant expressions and extract more semantically representative innovation concept units.

**Table 1.** Descriptive statistics of the preprocessing results.

| Metric                   | Mean    | Std.    | Min | Median | Max |
|--------------------------|---------|---------|-----|--------|-----|
| Raw tokenization tokens  | 217.626 | 146.751 | 37  | 182    | 870 |
| After stopword filtering | 139.759 | 94.387  | 27  | 117    | 574 |
| Final tokens             | 20.781  | 12.865  | 4   | 17     | 76  |

Furthermore, the gap between the minimum and maximum values suggests a clear long-tail distribution. Some teams produced concise idea texts, whereas others tended to adopt a “function list” description style, showing strong concept enumeration behavior. This heterogeneity is a typical feature of FFE idea expression. However, it may introduce bias in the accumulation of co-occurrence relations. To reduce the mechanical amplification effect caused by long texts and enumerative texts, this study adopted a document-based deduplication strategy in subsequent co-occurrence counting. That is, repeated co-occurrences of the same concept pair in the same idea text were counted only once. This ensures that edge weights reflect stable co-occurrence patterns across documents rather than repeated listing behavior in a single document.

#### 4.2. Co-Occurrence Network Construction

After concept extraction and structured representation of the idea texts, this study constructed a semantic co-occurrence network based on the sliding-window strategy in order to capture local semantic associations among innovation concepts in the idea texts. In the network, nodes represent concept tokens, edges represent the co-occurrence relationship of two tokens in a local window of the same text, and edge weights reflect the stability of co-occurrence across documents.

The results in Section 4.1 indicate that the idea texts vary substantially in scale. If an overly large window size is used, token pairs in short texts may form near-complete connections, leading to an overly dense network and large number of spurious co-occurrence. In contrast, an overly small window may weaken the ability to capture local semantic associations and may cause network fragmentation. Therefore, this study adopted a sliding-window strategy with window size  $S = 3$  and set the sliding step to 1, so as to achieve a balance between semantic locality and network connectivity.

To reduce the mechanical amplification of edge weights caused by repeated enumeration in a single idea text, a document-based deduplication counting strategy was applied. Specifically, each concept pair was counted at most once in each idea text. Edge weights were defined as the number of documents in which the concept pair co-occurred. This definition better reflects semantic associations across different idea texts and effectively reduces bias caused by long or enumerative descriptions.

Under the above rules and parameter settings, this study constructed a complete semantic co-occurrence network, which retained all co-occurrence relations observed within sliding windows, including weak links that appeared only once. The complete network contains 2100 nodes and 12,426 edges. The main reason for retaining weak links at this stage is that FFE idea texts exhibit strong heterogeneity and topic overlap. Weak edges often act as “semantic bridges” that maintain the overall connectivity of the semantic space and prevent excessive topic fragmentation. If all low-frequency edges were removed during network construction, the network could easily become fragmented, which would further affect the stability and interpretability of subsequent community detection.

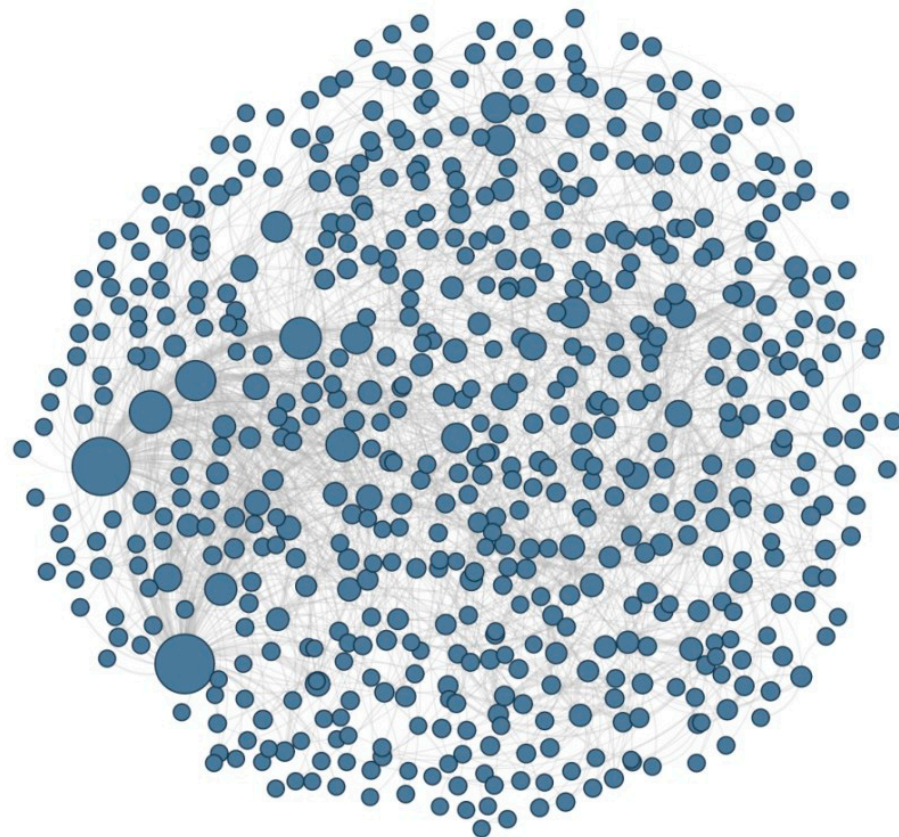
However, since subsequent measurements are sensitive to noisy edges, low-frequency co-occurrence relations may mainly reflect temporary expressions from individual teams rather than stable semantic structures shared across the idea texts. Therefore, to extract a representative semantic backbone network, this study introduced an edge weight threshold

during the backbone extraction stage. Only co-occurrence relations appearing in at least two idea texts were retained ( $\tau = 2$ ), thereby improving the robustness and interpretability of the measurement results.

The backbone network statistics show that the final semantic co-occurrence backbone network contains 598 nodes and 1604 edges. The network density is 0.0496, and the average node degree is 5.364, indicating that most tokens form stable co-occurrence relations with only a small number of other tokens. The network exhibits a sparse structure, which is consistent with the selective connectivity property commonly observed in semantic knowledge networks. Meanwhile, the maximum node degree reaches 91, indicating the existence of a small number of highly connected core concept nodes that act as hubs linking different semantic topics. Detailed network statistics are reported in Table 2. The backbone semantic network is illustrated in Figure 2.

**Table 2.** Descriptive statistics of the backbone network.

| Network Statistic     | Value  |
|-----------------------|--------|
| Number of nodes       | 598    |
| Number of edges       | 1604   |
| Network density       | 0.0496 |
| Average node degree   | 5.364  |
| Average node strength | 14.81  |
| Max node degree       | 91     |
| Max node strength     | 323    |



**Figure 2.** Backbone semantic network.

In terms of weighted connectivity strength, the average node strength is only 14.81, whereas the maximum strength reaches 323. This indicates a highly imbalanced distribution of co-occurrence relations. Specifically, a small number of concepts repeatedly appear in a

large number of idea texts and form stable combinations with various concepts, thereby constituting the high-frequency backbone of the semantic network. This result reflects the typical long-tail structural property of semantic networks that most concepts are low-frequency or local-topic nodes, while a small number of hub concepts form the main semantic structure shared across documents.

Further inspection of key hub nodes shows that high-strength nodes in the network mainly concentrate on two types of semantic elements: “demand-oriented concepts” and “industry core objects”. For example, “cultural demand” (Strength = 323) and “tea leaves” (Strength = 318) represent value propositions and core product elements, respectively, and both exhibit extremely high connectivity strength.

In addition, nodes such as “social demand”, “personalized demand”, and “health demand” also show relatively high strength. This suggests that innovation ideation in the corpus mainly revolves around themes such as social attributes, health value, and personalized experiences. To some extent, this distribution of hub concepts validates that the constructed semantic co-occurrence network is able to capture dominant value dimensions in the idea texts, providing a semantic basis for subsequent community-based measurements of cross-domain knowledge recombination indicator. Key hub nodes are shown in Table 3.

**Table 3.** Basic information on key hub nodes.

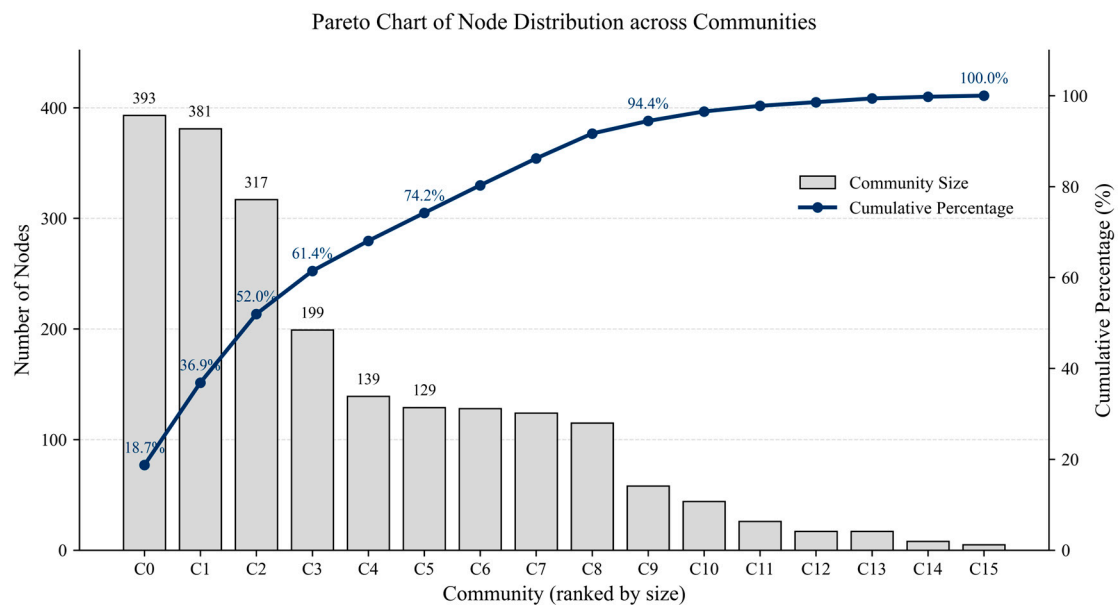
| Rank | Node                  | Degree | Strength |
|------|-----------------------|--------|----------|
| 1    | “cultural demand”     | 88     | 323      |
| 2    | “tea leaves”          | 91     | 318      |
| 3    | “social demand”       | 55     | 205      |
| 4    | “personalized demand” | 50     | 199      |

#### 4.3. Community Detection

After constructing the semantic co-occurrence network, community detection was conducted to extract the latent knowledge domain structure embedded in the idea corpus. Considering that innovation concepts in the FFE stage exhibit extensive cross-topic associations, and that weak co-occurrence edges may serve as topic-bridging connections in the semantic space, community detection was performed on the complete semantic co-occurrence network to preserve global connectivity.

Furthermore, as discussed in Section 4.2, the co-occurrence network exhibits typical sparsity and connection heterogeneity, this study applied the Leiden algorithm with a resolution parameter  $\gamma = 1$  to obtain an interpretable number of communities while avoiding overly fragmented partitions. After community detection, the number of nodes in each community is summarized in Figure 3.

The preliminary results indicate that the Leiden algorithm partitioned the complete semantic co-occurrence network into 16 co-occurrence communities, with an overall modularity of  $Q = 0.466$ . This suggests that the complete network exhibits a certain degree of modular structure. However, since the complete network retains a large number of weak co-occurrence edges with weight equal to 1, some communities may be structurally coupled through incidental connections. This may reduce modularity and cause certain communities to appear relatively dispersed in network visualization. Figure 4 shows the visualization of the detected communities in the semantic co-occurrence network.



**Figure 3.** Distribution of nodes across identified communities.



**Figure 4.** Community structure of the complete co-occurrence network.

Nevertheless, since the core measurement indicators proposed in this study require structurally robust semantic relations to reduce noise caused by incidental co-occurrence, this study further mapped the community labels obtained from the complete network onto the semantic backbone network. Modularity and community structural properties were then recalculated on the backbone network.

After mapping, the number of effective communities in the backbone network decreased to 12, and the overall modularity increased to  $Q = 0.619$ . This change indicates that after removing low-frequency noisy edges, the backbone network exhibits a clearer modular structure, with denser intra-community connections and sparser inter-community links. Consequently, the modularity optimization result becomes more stable and provides

stronger structural discriminability. Moreover, the modularity value of the backbone network serves as the baseline parameter for subsequent modularity change  $\Delta Q$  calculations.

#### 4.4. Measurement Results and Opportunity Ranking

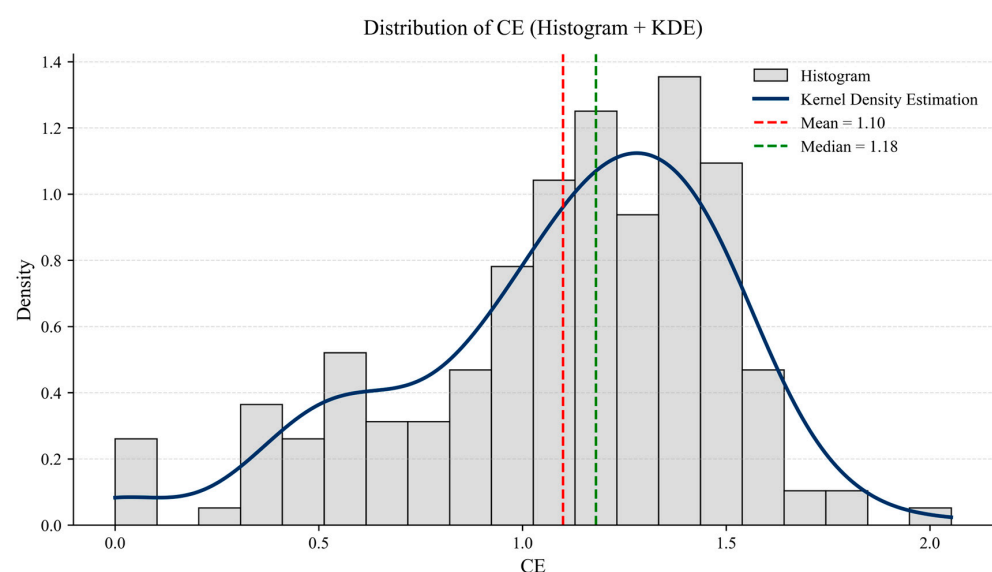
Based on the measurement framework proposed in Section 3, this study further calculated the cross-domain knowledge recombination indicator, the structural perturbation indicator, and the product innovation opportunity score for the 187 idea texts. All calculations were conducted based on the community partition results obtained from the backbone network in Section 4.3.

##### 4.4.1. Measurement Results of Cross-Domain Knowledge Recombination

The cross-domain knowledge recombination indicator  $CE$  is used to measure the extent to which a single idea text combines innovation concepts from different semantic communities. It reflects whether the idea breaks through a single knowledge domain and establishes cross-topic semantic connections.

The statistics of community proportion in each text show that each idea involves an average of 4.31 semantic communities, with the maximum spanning up to 10 communities. However, the mean value of the maximum community proportion reaches 0.541. This result suggests that although ideas often cover multiple community concepts, the majority of texts still display a combinatorial structure, with a primary knowledge domain supplemented by auxiliary knowledge domains. In other words, innovation ideas are typically centered around a primary thematic domain and are supplemented by concepts from other domains, rather than evenly integrating knowledge resources across multiple domains.

The computed cross-domain knowledge recombination indicator  $CE$  further demonstrates significant heterogeneity among idea texts. The mean value of the metric is 1.10. From the distribution perspective, most texts are concentrated in the range of 1.0–1.5, followed by 0.5–1.0, while relatively few texts exceed 1.5. This distribution pattern indicates that during the ideation process, most teams tend to expand their ideas by introducing a limited number of auxiliary-domain concepts based on a dominant domain. As a result, a common “moderate cross-domain fusion” pattern is formed. In contrast, ideas exhibiting a high level of cross-domain recombination are relatively rare, showing a clear long-tail characteristic. The rank order plot of  $CE$  is shown in Figure 5.



**Figure 5.** Distribution of  $CE$  across 187 idea texts.

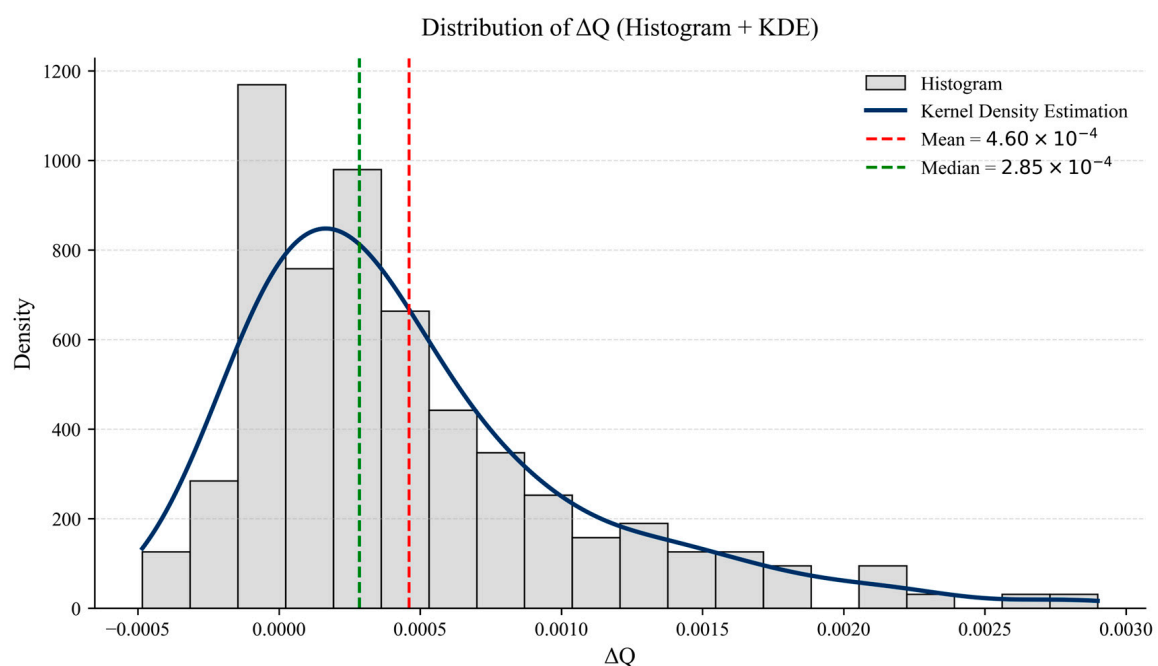
#### 4.4.2. Measurement Results of Structural Perturbation

Beyond the cross-domain knowledge recombination dimension, this study further measures the structural perturbation of idea texts from a network structure perspective, in order to capture the potential reconfiguration capability of their innovation concept combinations on the modular structure of the semantic network.

Unlike the cross-domain combinatorial entropy, which primarily reflects the degree of knowledge domain coverage and fusion, the structural perturbation metric focuses on whether the innovation concept connections involved in idea texts can significantly alter the community structure boundaries of the semantic network. That is, whether they possess the potential to break existing modular structures and introduce new cross-domain connection patterns.

Based on the structural perturbation measurement method proposed in Section 3, this study used the semantic backbone network obtained in Section 4.2 as the baseline structure, and adopted the modularity value  $Q_0 = 0.619$  derived from the backbone network community partition in Section 4.3 as the reference modularity parameter. For each idea text, the edges generated by that idea in the backbone network were removed. The modularity of the updated network  $Q_{-i}$  was then recalculated under the original community partition. The modularity change  $\Delta Q$  was computed for each idea, representing its structural perturbation degree.

The results show substantial heterogeneity in the structural perturbation dimension. The mean value of the structural perturbation metric is 0.00046. The distribution exhibits a typical right-skewed long-tail pattern. Most idea texts show extremely small structural perturbation. This implies that in the FFE stage, most ideas still follow the existing knowledge structure and mainly perform local extensions. Their innovation concept combinations have limited influence on the modular boundaries of the semantic network. In contrast, ideas with strong structural reconstruction potential are relatively scarce. The rank order plot of  $\Delta Q$  is shown in Figure 6.



**Figure 6.** Distribution of  $\Delta Q$  across 187 idea texts.

It is noteworthy that the maximum value of the structural perturbation metric is 0.0028, while the minimum value reaches  $-0.00048$ . This indicates that the indicator contains not only positive samples but also a certain proportion of negative samples. This phenomenon

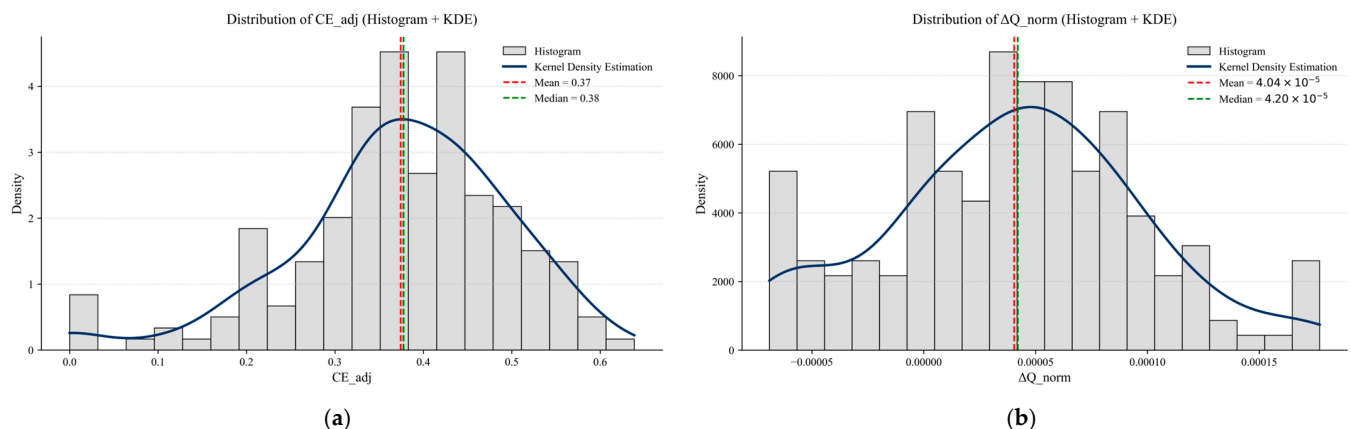
suggests that the influence of idea texts on the modular structure of the semantic network is not limited to a single direction. Instead, two structurally opposite mechanisms may coexist. Therefore, the sign of  $\Delta Q$  reflects not only the strength of perturbation but also the direction of the idea's structural impact on the network.

#### 4.4.3. Measurement Results of Product Innovation Opportunity Score

To integrate both signals from cross-domain knowledge recombination and structural perturbation, this study further computed the *PIOS* based on the composite mechanism proposed in Section 3. The *PIOS* was calculated for all 187 idea texts. By ranking this metric, early-stage product innovation opportunities can be assessed in a unified framework. *PIOS* is able to capture both the breadth of knowledge recombination and the structural reconstruction potential within a unified framework, thereby reducing bias caused by reliance on a single indicator.

Before calculating the composite *PIOS*, the two metrics (*CE* and  $\Delta Q$ ) were first adjusted according to their respective structural characteristics. Specifically, the cross-domain knowledge recombination indicator was transformed into a logarithmically adjusted entropy measure ( $CE_{i,adj}$ ) to alleviate potential inflation caused by excessively long or enumeration-style ideas. Meanwhile, the structural perturbation indicator was normalized into  $\Delta Q_{i,norm}$  by dividing the original perturbation value by the structural participation strength of each idea, thereby reducing the influence of structural participation scale on perturbation magnitude.

Figure 7 presents the distributions of the adjusted structural metrics. Overall, both adjusted metrics exhibit relatively stable and concentrated distribution patterns, suggesting that the proposed adjustment strategies effectively alleviate potential inflation effects caused by token quantity and structural participation scale while preserving the underlying structural characteristics of the original metrics.



**Figure 7.** Distribution of the adjusted structural metrics: (a) Adjusted cross-domain knowledge recombination indicator ( $CE_{i,adj}$ ); (b) Normalized structural perturbation indicator ( $\Delta Q_{i,norm}$ ).

As shown in Figure 7a, compared with the original *CE* distribution, the adjusted cross-domain knowledge recombination metric ( $CE_{i,adj}$ ) exhibits a noticeably alleviated long-tail characteristic while maintaining the overall unimodal distribution pattern. This suggests that the logarithmic adjustment effectively moderates the inflation effect caused by excessively long or enumeration-style ideas without substantially distorting the overall distribution structure.

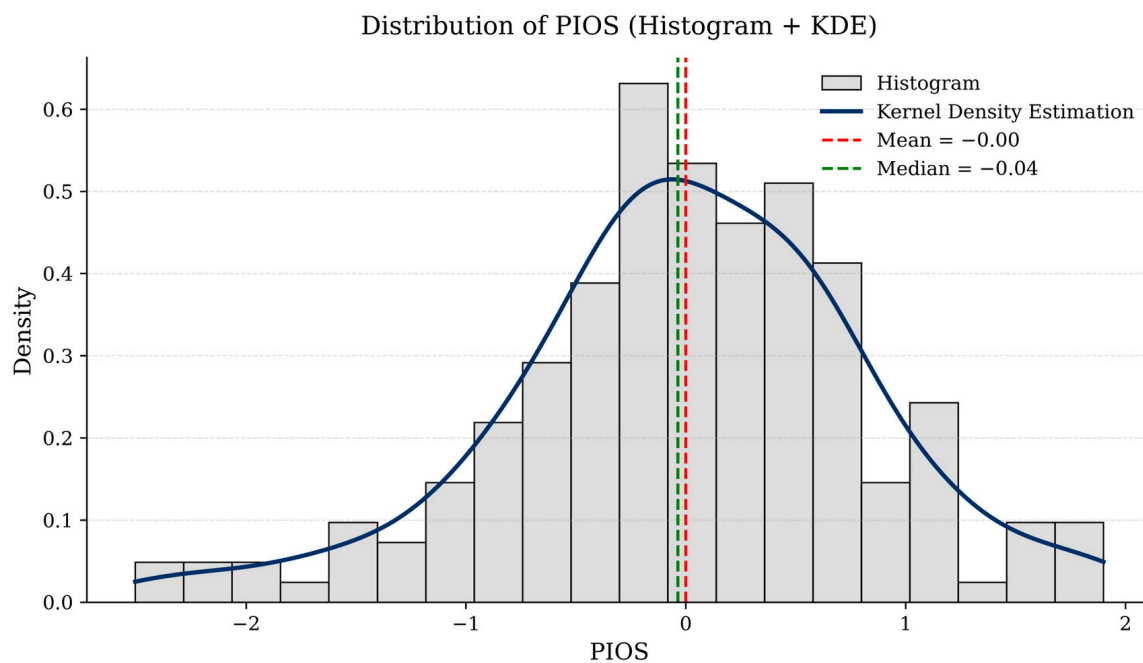
Figure 7b shows the distribution of the normalized structural perturbation metric ( $\Delta Q_{i,norm}$ ). Compared with the original  $\Delta Q$  distribution, the normalized metric becomes more concentrated and balanced, with highly similar mean and median values. Although several relatively high perturbation observations are still preserved in the right tail, the

overall distribution becomes less dominated by structural participation scale differences across idea texts. This indicates that the proposed normalization effectively moderates perturbation inflation while retaining highly efficient structural perturbation signals.

Since the adjusted  $CE_{i,adj}$  and normalized  $\Delta Q_{i,norm}$  metrics still differed significantly in numerical scale and magnitude, both metrics were further standardized using Z-score transformation before  $PIOS$  calculation. This ensures comparability across dimensions and prevents one metric from dominating the results due to scale differences.

After standardization, the cross-domain knowledge recombination degree  $CE$  and the structural perturbation degree  $\Delta Q$  were integrated based on the  $PIOS$  calculation procedure proposed in Section 3.5.3. A unified opportunity score was thus obtained, enabling the final ranking of all idea texts.

The rank order plot of  $PIOS$  is shown in Figure 8. The  $PIOS$  exhibit a clear non-uniform distribution pattern. Most ideas are concentrated around the mean, mainly within the range of approximately  $-1$  to  $1$ , indicating that the majority of ideas show relatively limited differences in product innovation opportunity characteristics. However, clear tail structures appear at both ends of the distribution. On the one hand, a small number of  $PIOS$  significantly above the mean and form an obvious right-sided long tail, with a maximum value of  $1.90$ . This suggests that certain ideas demonstrate outstanding performance in both cross-domain knowledge integration and structural reconstruction. On the other hand, a small number of ideas score significantly below the mean and form a left-sided tail, with a minimum value of approximately  $-2.50$ . This indicates that some ideas tend to follow a localization deepening pattern or a structure-stabilizing recombination mode. Overall, this result suggests that high-potential product innovation opportunities are significantly scarce in the dataset. High-potential opportunities are mainly concentrated in a small number of idea texts rather than being evenly distributed across all ideas.



**Figure 8.** Distribution of  $PIOS$  across 187 idea texts.

Further examination of the top-ranked ideas shows that these ideas often simultaneously exhibit high cross-domain recombination entropy and strong positive structural perturbation effects. Their innovation concept combinations not only cover multiple knowledge communities but also form stable cross-domain bridging structures in the semantic

backbone network, thereby weakening the existing modular boundaries and introducing new knowledge coupling patterns, leading to the disruption of community boundaries.

In addition, some top-ranked ideas show a relatively low value in one metric but an exceptionally high value in the other. These cases further demonstrate the complementarity between cross-domain knowledge recombination and structural perturbation in product innovation opportunity identification. The *PIOS* effectively integrates these complementary signals, preventing potentially valuable opportunities from being overlooked simply because one indicator is slightly lower. The top 10 ideas ranked by *PIOS* are summarized in Table 4.

**Table 4.** The top 10 ideas ranked by *PIOS*.

| $i$ | $PIOS_i$ | $CE'_i$ | $\Delta Q'_i$ |
|-----|----------|---------|---------------|
| 178 | 1.900    | 1.864   | 1.936         |
| 182 | 1.888    | 1.588   | 2.189         |
| 21  | 1.850    | 2.179   | 1.520         |
| 25  | 1.841    | 1.467   | 2.216         |
| 39  | 1.646    | 1.786   | 1.506         |
| 73  | 1.612    | 0.816   | 2.409         |
| 187 | 1.580    | 1.651   | 1.509         |
| 18  | 1.505    | 1.786   | 1.224         |
| 74  | 1.291    | 1.200   | 1.382         |
| 183 | 1.239    | 0.476   | 2.001         |

In summary, based on the proposed measurement framework, this study ultimately identified the most likely product innovation opportunities in idea texts related to tea products and tea-related services.

## 5. Validation

To further evaluate the effectiveness and robustness of the proposed measurement framework, this section conducts criterion validity analysis, weighting sensitivity analysis, and comparative case analysis. These analyses provide empirical support for the reliability and interpretability of the proposed framework.

### 5.1. Criterion Validity

To examine whether the proposed measurement framework has practical applicability and whether it is consistent with human expert judgments, this study conducted a criterion-related validity test. According to measurement theory, criterion validity is an important method for evaluating whether a newly developed measurement can reflect the target construct. It is typically assessed by examining the correlation between the proposed measure and an external criterion variable.

Therefore, this study adopted expert evaluation as an external criterion to validate the framework. Specifically, five experts with relevant domain experience were invited to evaluate the innovation potential of the selected ideas in an early-stage product innovation sense. Experts were instructed to assess whether the focal idea showed promise as a product or service innovation concept in terms of originality, relevance, and developmental potential, rather than whether later commercial success had already occurred. The evaluation used three items on a five-point Likert scale ranging from 1 to 5, where higher scores indicate stronger product innovation opportunity potential. An example item is: “To what extent does the new product/service proposed by this team represent a promising product innovation opportunity?”

To examine inter-rater consistency among experts, the intraclass correlation coefficient (ICC) was calculated. The results show high agreement among experts (ICC = 0.81), indicating strong consistency in experts' evaluations of product innovation opportunities. Therefore, the mean value of the five experts' scores was used as the final expert evaluation score.

To enhance discriminability between different levels of product innovation opportunity, this study further adopted an extreme groups design. Based on the rank of *PIOS*, the 15 highest-scoring and 15 lowest-scoring ideas were selected, resulting in a total of 30 samples for expert evaluation. This strategy improves the identifiability of differences across samples.

The correlation analysis results show that the framework is significantly positively correlated with expert evaluation scores ( $r = 0.389, p < 0.05$ ). This indicates that the proposed measurement framework is able to effectively reflect informed early-stage judgments of product innovation potential, thereby supporting its criterion-related validity. The results of the correlation analysis are shown in Table 5.

**Table 5.** Correlation between *PIOS* and expert evaluation scores ( $N = 30$ ).

| Variable                                  | Product Innovation Opportunity Indicators | Expert Rating |
|-------------------------------------------|-------------------------------------------|---------------|
| Product Innovation Opportunity Indicators | 1                                         |               |
| Expert Rating                             | 0.389 *                                   | 1             |

Note: \*  $p < 0.05$  (two-tailed test).

Overall, this framework, constructed based on the semantic co-occurrence network, not only captures the cross-domain recombination characteristics of knowledge elements from a structural perspective but also shows good empirical consistency with expert subjective evaluations of early-stage product innovation potential. The criterion validity results provide preliminary empirical support for the practical effectiveness and interpretability of the proposed framework.

## 5.2. Weighting Robustness

Since the proposed measurement framework adopts a linear aggregation strategy with equal weighting between the cross-domain recombination dimension and the structural perturbation dimension, additional sensitivity analysis was conducted to examine whether different weighting configurations would substantially affect the final ranking results. This analysis aims to evaluate the robustness of the equal-weight aggregation strategy and determine whether moderate variations in weighting allocation would significantly alter the identification results of early-stage product innovation opportunities.

Specifically, the weighting parameter  $\alpha$  in the *PIOS* aggregation formula was systematically adjusted from 0.1 to 0.9 at intervals of 0.1, where larger values indicate greater emphasis on cross-domain knowledge recombination and smaller values indicate greater emphasis on structural perturbation. The equal-weight configuration ( $\alpha = 0.5$ ) was treated as the baseline condition. Spearman rank correlation analysis and Top-10 overlap analysis were subsequently conducted to evaluate the stability of the ranking structure under different weighting schemes.

The Spearman rank correlation results are presented in Table 6. The results indicate that the ranking outcomes under different weighting configurations remained highly consistent with the baseline equal-weight setting. Specifically, all weighting schemes produced correlation coefficients above 0.80, while most configurations yielded coefficients exceeding 0.90. The ranking consistency gradually increased as the weighting configuration

approached the equal-weight condition, suggesting that moderate variations in weighting allocation do not substantially alter the overall ranking structure generated by the proposed framework.

**Table 6.** Sensitivity Analysis of Ranking Stability Under Different Weighting Configurations.

| Weight Configuration | Spearman $\rho$ with Baseline (0.5:0.5) | Stability Interpretation |
|----------------------|-----------------------------------------|--------------------------|
| 0.1:0.9              | 0.8655                                  | Moderate consistency     |
| 0.2:0.8              | 0.9133                                  | Strong consistency       |
| 0.3:0.7              | 0.9559                                  | Strong consistency       |
| 0.4:0.6              | 0.9871                                  | Very strong consistency  |
| 0.5:0.5              | 1.0000                                  | Baseline                 |
| 0.6:0.4              | 0.9853                                  | Very strong consistency  |
| 0.7:0.3              | 0.9462                                  | Strong consistency       |
| 0.8:0.2              | 0.8884                                  | Moderate consistency     |
| 0.9:0.1              | 0.8181                                  | Moderate consistency     |

To further evaluate the robustness of the highest-ranked opportunity ideas, a Top-10 overlap analysis was additionally conducted, as shown in Table 7.

**Table 7.** Top-10 Overlap Analysis Under Different Weighting Configurations.

| Weight Configuration | Top-10 Overlap | Overlap Ratio | Stability Interpretation |
|----------------------|----------------|---------------|--------------------------|
| 0.1:0.9              | 6              | 60%           | Moderate                 |
| 0.2:0.8              | 7              | 70%           | Relatively robust        |
| 0.3:0.7              | 8              | 80%           | Relatively robust        |
| 0.4:0.6              | 9              | 90%           | Highly robust            |
| 0.5:0.5              | 10             | 100%          | Baseline                 |
| 0.6:0.4              | 9              | 90%           | Highly robust            |
| 0.7:0.3              | 9              | 90%           | Highly robust            |
| 0.8:0.2              | 8              | 80%           | Relatively robust        |
| 0.9:0.1              | 7              | 70%           | Relatively robust        |

The results demonstrate that the majority of Top-10 ideas identified under the equal-weight setting were consistently retained across alternative weighting configurations. In most cases, the overlap ratio ranged from 70% to 90%, while adjacent weighting configurations achieved particularly high overlap stability. Although several opportunity ideas entered or exited the Top-10 list under more extreme weighting conditions, the overall high-potential opportunity set remained relatively stable.

Overall, the sensitivity analysis demonstrates that the proposed measurement framework maintains relatively stable ranking behavior under different weighting configurations. Therefore, under conditions where no established theoretical or empirical evidence supports differential weighting between the two dimensions, the equal-weight linear aggregation strategy adopted in this study can be regarded as a transparent, interpretable, and robust methodological choice.

### 5.3. Comparative Case Analysis

A recurring challenge in text-based innovation opportunity identification is that semantic novelty does not necessarily imply a strong product innovation opportunity. Ideas that merely combine heterogeneous concepts may exhibit high cross-domain recombination characteristics while still failing to reshape the underlying semantic structure of the industry. To further illustrate the discriminative capability of the proposed two-dimensional

framework, this section compares representative idea cases with different combinations of cross-domain recombination and structural perturbation characteristics.

A representative example of high cross-domain recombination but low structural perturbation is Idea FNo164 (“Cozy House”, Rank: 114). This idea integrates multiple heterogeneous concepts, including massage equipment, tea aroma humidifiers, biodegradable technology, projection systems, and wellness-oriented experience functions. Such combinations significantly increase the semantic diversity and cross-domain connectivity of the idea, thereby contributing to a relatively high cross-domain recombination value ( $CEI = 1.3679$ ).

However, despite the presence of multiple cross-domain elements, the idea mainly represents an additive integration of existing functional concepts rather than a structural reconfiguration of the underlying semantic relationships within the tea industry network. Most semantic associations introduced by the idea remain peripheral extensions surrounding existing consumption and wellness concepts, without substantially altering the original semantic community structure of the industry network ( $\Delta QI = -1.6678$ ). Therefore, although the idea demonstrates relatively strong semantic novelty, its structural perturbation effect remains limited. As a result, the idea was ranked only 114th under the framework despite its relatively rich semantic diversity.

In contrast, Idea FNo187 (“Self-service Tea Machine”, Rank: 7) achieved high scores in both cross-domain recombination and structural perturbation ( $CEI = 1.6514$ ,  $\Delta QI = 1.5091$ ). The idea integrates concepts such as intelligent vending systems, sensors, application-based control technologies, and automated tea preparation mechanisms. More importantly, the idea reconstructs the traditional semantic relationships associated with tea consumption scenarios.

Specifically, the idea shifts tea consumption from manually prepared and location-dependent experiences toward automated, instant, and scenario-flexible consumption patterns. This restructuring effect introduces new semantic associations between tea products, intelligent devices, convenience services, and personalized consumption contexts, thereby generating substantial perturbation to the original semantic community structure. As a result, the idea demonstrates not only semantic novelty but also stronger restructuring potential within the semantic network. Ultimately, it ranked in the top-10 within the framework.

These comparative cases further demonstrate that cross-domain recombination alone is insufficient for assessing product innovation opportunities. While semantically novel ideas may introduce heterogeneous concept combinations, they do not necessarily reshape the underlying semantic organization of the idea space. By incorporating structural perturbation analysis, the proposed framework is able to distinguish merely novel semantic combinations from ideas with stronger restructuring potential. This further highlights the necessity of integrating semantic recombination analysis with structural perturbation analysis when assessing early-stage product innovation opportunities from text-based idea descriptions.

## 6. Discussion

This study develops a text co-occurrence-network-based decision-support system for assessing early-stage product innovation opportunities in the fuzzy front end of new product development. Its central theoretical implication is that the assessment of early-stage product innovation opportunities can be moved from a retrospective logic based on realized outcomes to an ex ante logic based on semantic structure and knowledge-system positioning. Because firms must make screening and resource-allocation decisions before

patent citations, market diffusion, or performance outcomes become available [1,3,7,8], such a shift is especially important for front-end innovation governance.

From the perspective of knowledge recombination theory, innovation can be understood as a process of reorganizing existing knowledge structures into new configurations [10–16]. Under this theoretical lens, idea texts are not merely descriptive narratives of product concepts; rather, they represent one of the few accessible information sources that contain early evidence of recombination activities in the FFE stage [3,4,26]. Accordingly, this study treats idea texts as structural evidence of early-stage recombination. This provides a feasible pathway for screening ideas that deviate from mainstream trajectories and exhibit stronger product innovation potential while recognizing that later commercial outcomes still depend on downstream development and market testing.

Importantly, product innovation potential should not be interpreted simply as the presence of novel or rare concepts. Instead, it should be interpreted as the way in which concepts are integrated across knowledge domains and how those integrations suggest new product meanings, use scenarios, or value propositions. This means evaluating not only semantic novelty but also whether an idea opens plausible new directions for products or services. Many text-based early indicators emphasize the use of rare words, semantic distance, or token uniqueness [22,28,30,31]. While these indicators may reflect distinctiveness, they may not distinguish structurally meaningful recombination from ideas that are merely unusual in wording. In contrast, this study uses a semantic co-occurrence network to represent recombination behavior, where co-occurrence links reflect observed coupling relationships in idea texts [7,8]. Compared with token novelty measures, this representation enables product innovation potential to be examined through recombination patterns rather than isolated term attributes.

Within knowledge recombination theory, ideas with stronger product innovation potential are more likely to emerge from the integration of concept sets that are cognitively distant, and such integration should generate structural impacts on the existing knowledge landscape [12–16]. Therefore, the proposed framework measures product innovation opportunities through two complementary dimensions of recombination. First, cross-domain recombination reflects whether an idea draws innovation concepts from multiple semantic communities. This is measured by community entropy, where a higher entropy value indicates a broader search scope and a greater likelihood of unconventional integration. However, broad integration alone does not necessarily guarantee structural change. Second, structural perturbation captures whether the associations contributed by an idea are likely to influence the modular separation of the semantic network. This dimension directly complements the limitation of cross-domain recombination by reflecting potential structural reconfiguration effects [11,16,35,49].

The construction of *PIOS* adopts a linear aggregation strategy to integrate cross-domain knowledge recombination and network structural perturbation. This strategy treats the two dimensions as separable structural signals contributing to the overall score in an additive form. In contrast, nonlinear aggregation forms such as geometric combinations impose stronger coupling constraints between dimensions, which may be overly restrictive for exploratory evaluation settings. Furthermore, the sensitivity analysis reported in Section 5 suggests that the overall ranking pattern is stable under alternative weighting configurations.

The observed consistency between model outputs and expert evaluations further indicates that the proposed indicators align with domain-relevant cognitive judgments of early-stage product innovation potential rather than merely reflecting structural artifacts of text co-occurrence networks.

Moreover, these two dimensions are not equivalent but mutually complementary. An idea may involve multiple communities but exhibit limited integration depth. Conversely, another idea may involve fewer communities but introduce a small number of structurally meaningful bridging connections. This distinction suggests that interpreting product innovation opportunities requires considering both the breadth of recombination and the degree of structural impact in the knowledge system [20,50–52]. It also reinforces the need to distinguish between early-stage opportunities and realized outcomes.

This interpretation is further supported by the counterexample identified in the Validation Section, where an idea exhibits high cross-domain recombination (*CE*) but low structural perturbation ( $\Delta Q$ ). Such cases reflect superficial semantic juxtaposition across distant domains without inducing substantial reconfiguration of the underlying knowledge network structure. This evidence reinforces the necessity of jointly considering both dimensions, as *CE* alone cannot fully capture whether cross-domain combinations translate into structurally meaningful product innovation opportunities in the early-stage fuzzy front end of new product development within the knowledge system.

Nevertheless, relying solely on idea texts is insufficient because recombination is only meaningful when evaluated within an established and realistic knowledge environment. If the semantic network is constructed only from experimental idea texts, the resulting structure may deviate from the actual market knowledge landscape and fail to reflect the real industry-level knowledge system. In such cases, the network would have limited practical relevance. By incorporating market corpus texts, the semantic network can capture a broader and more stable knowledge structure. This helps distinguish ideas that are structurally anomalous relative to mainstream semantic associations from those that only appear anomalous within a limited experimental dataset.

At the same time, the effectiveness of the external market corpus is inherently time-sensitive because market discourse, consumer preferences, and industry terminology may evolve over time. Therefore, in practical enterprise applications, the external corpus should be periodically updated according to the semantic evolution speed of the target domain. For relatively stable consumer-oriented industries such as the tea industry examined in this study, semantic structures may remain stable over longer periods, whereas technology-intensive industries may require more frequent corpus updating to maintain the validity of semantic boundary detection.

The proposed framework is not inherently restricted to the tea industry, because its analytical logic is based on general semantic network mechanisms, including concept co-occurrence, community structure, cross-domain recombination, and structural perturbation. Therefore, the framework can theoretically be transferred to other industrial contexts where idea descriptions and external market knowledge are available. However, the adaptation process requires adjustments according to the characteristics of the target domain. In particular, the external market corpus should reflect the dominant knowledge structure and communication style of the industry under investigation. For example, consumer-oriented industries may rely more on user reviews, social media discussions, and product descriptions, whereas technology-intensive industries may require patents, technical reports, academic abstracts, or developer documentation as external semantic references.

In addition, different industrial corpora may require different preprocessing and network-construction strategies due to variations in terminology density, document structure, and semantic stability. Technical or patent-like texts, for instance, may benefit from larger co-occurrence windows, domain-specific synonym expansion, or protected technical-term dictionaries to preserve semantic integrity. Moreover, because industry terminology and market knowledge evolve over time, the external corpus should be periodically up-

dated in practical applications to maintain the validity of the semantic network structure and avoid outdated semantic boundaries.

Furthermore, a key methodological consideration in this study is that semantic and structural signals extracted from FFE idea texts may be systematically influenced by surface-level expression patterns, particularly variation in text length and enumeration-style descriptions. Such biases are common in early-stage ideation data, where some teams tend to list multiple loosely related concepts while others describe ideas more concisely. To address this issue, this study applies bias-adjustment strategies prior to aggregation. Specifically, logarithmic length adjustment is introduced for the cross-domain recombination measure, and structural perturbation is normalized by structural participation strength. These procedures ensure that both indicators reflect underlying semantic and structural heterogeneity rather than artefacts of expression scale.

From a managerial perspective, the proposed framework can be understood as a screening mechanism embedded in enterprise innovation systems. In practice, the FFE process often generates a large number of heterogeneous ideas, while evaluation resources are limited and costly [1,5,6,53]. The text-network indicators developed here provide a scalable means of prioritizing ideas that show stronger signals of cross-domain recombination and boundary-spanning structure, thereby improving opportunity recognition, portfolio selection, and the operational efficiency of front-end review processes. Accordingly, the framework contributes not only to innovation theory but also to the design of decision-support systems for early-stage innovation governance.

These indicators should not be interpreted as deterministic predictors of eventual commercial success. Rather, they provide early signals that can guide managerial attention, expert review, and subsequent experimentation under uncertainty. The framework should therefore be viewed as a socio-technical decision aid that complements, rather than replaces, human evaluation. This positioning is especially important because the eventual realization of product innovation still depends on downstream technical feasibility, adoption dynamics, organizational execution, and competitive response. The framework thus addresses a specific but critical problem: how to improve the quality and consistency of front-end screening when reliable outcome-based evidence does not yet exist.

Despite these contributions, the proposed framework also has limitations. First, the data object of this framework is unstructured text, meaning that it requires a certain level of text quality and completeness. Brief, incomplete, or vague descriptions may lead to lower-than-expected *PIOS*, whereas overly long and enumerative descriptions may produce inflated *PIOS*. Second, co-occurrence relationships remain a relatively simple contextual association and do not capture causal or functional relationships between technologies [27]. Third, because the current framework emphasizes semantic-structural signals, it does not yet directly model several market-side and demand-side cues that may matter for product innovation assessment, such as user needs, willingness to pay, adoption frictions, or commercialization constraints. Fourth, this study does not explore the optimal weighting of the two dimensions in *PIOS*. A linear aggregation strategy is adopted, but alternative weighting schemes may yield improved performance. Finally, although the semantic network construction and detection procedures enable lightweight screening, additional efforts are required to build domain-specific resources, such as protected terms, stopwords, synonym mappings, and market corpora. Moreover, because the framework depends on parameter settings, sensitivity analyses across industries and corpora are necessary to ensure model stability. This also means that the current framework is not yet suitable for evaluating technology commercialization opportunities, technology transfer pathways, or licensing-oriented commercialization decisions associated with relatively mature technological assets. Accordingly, the present study should be understood as a

front-end screening framework for product and service idea texts, rather than as a direct assessment tool for the commercialization potential of technologies.

## 7. Conclusions and Future Research

This study proposes a text co-occurrence-network-based decision-support framework for assessing early-stage product innovation opportunities in the fuzzy front end of new product development. By shifting attention from lagged outcome indicators to ex ante semantic-structural evidence, the framework offers a practical way to evaluate early ideas when citations, market signals, and diffusion traces are not yet available [1,3,8,49,54]. Grounded in knowledge recombination theory [10–16], the approach converts unstructured idea texts into measurable semantic relations through concept extraction, co-occurrence network construction, community detection, and metric computation.

The contribution of this study lies in translating knowledge recombination mechanisms into quantifiable semantic signals and enabling the assessment of early-stage product innovation opportunities from an ex ante perspective. The proposed framework is reproducible, transparent, and scalable. It does not rely on lagged outcome indicators and can reduce exclusive dependence on subjective expert judgment in the FFE [8,53]. More broadly, the study contributes to research on product innovation assessment by showing that early opportunity potential can be approached as an idea-level structural property, rather than only as a realized outcome visible after diffusion or launch.

In practical terms, the proposed framework can be embedded into FFE innovation governance as an intelligent screening and prioritization module. Organizations often face large volumes of heterogeneous idea submissions while having limited expert evaluation capacity [1,5,6]. The framework therefore provides a scalable way to identify structurally anomalous and potentially boundary-spanning ideas for further review, supporting portfolio-level decision making and resource allocation under uncertainty. In this respect, the framework is especially relevant for firms seeking to strengthen the operational efficiency and analytical consistency of front-end innovation processes.

Future research may extend this framework in several directions. First, semantic similarity networks can be combined with co-occurrence networks to improve the representation of relationships between idea tokens. For example, similarity-based networks may facilitate synonym mapping and reduce missing semantic links, thereby enhancing interpretability [26,27]. Second, the quantitative evaluation metrics for FFE ideas can be further optimized, such as by improving the weighting strategy between the two dimensions or introducing additional dimensions to provide a more comprehensive product innovation opportunity assessment. Third, future work should incorporate market-side and demand-side cues more directly, including user needs, value propositions, adoption frictions, and commercialization constraints that may shape early opportunity evaluation. Fourth, the proposed model should be validated across more industries to establish a more generalizable measurement framework. Fifth, a deeper investigation of community structural mechanisms is needed, such as whether different types of ideas produce different directions or patterns of influence on community structures [26,27,31]. Sixth, future research could extend the knowledge recombination perspective to newly submitted patent texts and related technology documents. Such materials may provide a more appropriate empirical basis for examining technology commercialization opportunities and for assessing the innovation potential of technologies themselves.

Overall, this study proposes a semantic co-occurrence-network-based measurement framework for assessing early-stage product innovation opportunities in the FFE stage. The framework should be understood as an interpretable decision-support system that facilitates early-stage screening and attention allocation under uncertainty while recognizing

ing that eventual success still depends on downstream feasibility, adoption dynamics, and organizational execution. Its main value lies in moving opportunity assessment closer to the point where firms actually make front-end decisions, thereby helping bridge the gap between innovation theory and early-stage innovation practice.

**Author Contributions:** Conceptualization, Z.W. and G.Q.; Methodology, Z.W. and G.Q.; Software, S.G. and P.L.; Validation, G.Q.; Formal analysis, S.G.; Resources, D.H.; Data curation, P.L.; Writing—original draft, Z.W., S.G. and P.L.; Writing—review and editing, G.Q. and D.H.; Visualization, S.G.; Supervision, G.Q. and D.H.; Project administration, Z.W. and D.H.; Funding acquisition, Z.W. and D.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Social Science Fund of China (grant number 21BGL067).

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Review Committee of School of Economics and Management, Fuzhou University (protocol code REA211021-03 and date of approval 15 October 2023).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. Public release of the dataset is restricted because the corpus contains original idea-generation materials produced by research participants, which may involve unpublished innovation concepts and potential intellectual property considerations.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Khurana, A.; Rosenthal, S.R. Integrating the Fuzzy Front End of New Product Development. *MIT Sloan Manag. Rev.* **1997**, *38*, 103–120.
2. Cooper, R. Perspective: The Stage-Gate® Idea-to-Launch Process—Update, What’s New, and NexGen Systems. *J. Prod. Innov. Manag.* **2008**, *25*, 213–232. [[CrossRef](#)]
3. Koen, P.; Ajamian, G.; Burkart, R.; Clamen, A.; Davidson, J.; D’Amore, R.; Elkins, C.; Herald, K.; Incorvia, M.; Johnson, A.; et al. Providing clarity and a common language to the “fuzzy front end”. *Res. Technol. Manag.* **2001**, *44*, 46–55. [[CrossRef](#)]
4. Kim, J.; Wilemon, D. Focusing the Fuzzy Front-End in New Product Development. *R Manag.* **2003**, *32*, 269–279. [[CrossRef](#)]
5. Reid, S.; de Brentani, U. The Fuzzy Front End of New Product Development for Discontinuous Innovations: A Theoretical Model. *J. Prod. Innov. Manag.* **2004**, *21*, 170–184. [[CrossRef](#)]
6. Girotra, K.; Terwiesch, C.; Ulrich, K.T. Idea Generation and the Quality of the Best Idea. *Manag. Sci.* **2010**, *56*, 591–605. [[CrossRef](#)]
7. Verhoeven, D.; Bakker, J.; Veugelers, R. Measuring Technological Novelty with Patent-Based Indicators. *Res. Policy* **2016**, *45*, 707–723. [[CrossRef](#)]
8. Funk, R.J.; Owen-Smith, J. A Dynamic Network Measure of Technological Change. *Manag. Sci.* **2017**, *63*, 791–817. [[CrossRef](#)]
9. Schumpeter, J.A. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*; Transaction Publishers: Piscataway, NJ, USA, 1934.
10. Morgenstern, O.; Schumpeter, J.A. Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process. *J. Am. Stat. Assoc.* **1940**, *35*, 423. [[CrossRef](#)]
11. Fleming, L. Recombinant Uncertainty in Technological Search. *Manag. Sci.* **2001**, *47*, 117–132. [[CrossRef](#)]
12. Rosenkopf, L.; Nerkar, A. Beyond Local Search: Boundary-Spanning, Exploration, and Impact in the Optical Disk Industry. *Strateg. Manag. J.* **2001**, *22*, 287–306. [[CrossRef](#)]
13. Ahuja, G.; Morris Lampert, C. Entrepreneurship in the Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Inventions. *Strateg. Manag. J.* **2001**, *22*, 521–543. [[CrossRef](#)]
14. Katila, R.; Ahuja, G. Something Old, Something New: A Longitudinal Study of Search Behavior and New Product Introduction. *Acad. Manag. J.* **2002**, *45*, 1183–1194. [[CrossRef](#)]
15. Uzzi, B.; Mukherjee, S.; Stringer, M.; Jones, B. Atypical Combinations and Scientific Impact. *Science* **2013**, *342*, 468–472. [[CrossRef](#)] [[PubMed](#)]
16. Reinertsen, D.G. Taking the Fuzziness Out of the Fuzzy Front End. *Res. Technol. Manag.* **1999**, *42*, 25–31. [[CrossRef](#)]

17. Gåsvaer, D.; Fundin, A.; Johansson, P.; Langbeck, B. Mind the Gap—Managing Boundaries in the Fuzzy Front End of Production Innovation. *Eur. J. Innov. Manag.* **2025**, *28*, 187–208. [\[CrossRef\]](#)
18. Kurt, O.E.; Vayvay, O. Gravitational Intelligent Decision-Making Model at the Fuzzy Front End with Extrinsic Idea Integration by the K-Means Algorithm. *Systems* **2022**, *10*, 194. [\[CrossRef\]](#)
19. Ma, M.; Mao, J.; Li, G. Discovering Weak Signals of Emerging Topics with a Triple-Dimensional Framework. *Inf. Process. Manag.* **2024**, *61*, 103793. [\[CrossRef\]](#)
20. Ebadi, A.; Auger, A.; Gauthier, Y. Detecting Emerging Technologies and Their Evolution Using Deep Learning and Weak Signal Analysis. *J. Informetr.* **2022**, *16*, 101344. [\[CrossRef\]](#)
21. Zheng, L.; Sun, L.; He, Z.; He, S. The Identification of Dynamic Product Innovation Opportunities Using the Multi-Phase QFD: The Customer Requirement and Technology Development Perspectives. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 1351. [\[CrossRef\]](#)
22. Antons, D.; Grünwald, E.; Cichy, P.; Salge, O. The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities. *RD Manag.* **2020**, *50*, 329–351. [\[CrossRef\]](#)
23. Callon, M.; Courtial, J.P.; Laville, F. Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemsitry. *Scientometrics* **1991**, *22*, 155–205. [\[CrossRef\]](#)
24. Van Eck, N.J.; Waltman, L. Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. *Scientometrics* **2010**, *84*, 523–538. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Boyack, K.W.; Klavans, R. Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2389–2404. [\[CrossRef\]](#)
26. Traag, V.; Waltman, L.; Eck, N.J. van From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* **2019**, *9*, 5233. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Newman, M.E.J.; Girvan, M. Finding and Evaluating Community Structure in Networks. *Phys. Rev. E* **2004**, *69*, 026113. [\[CrossRef\]](#)
28. Arts, S.; Cassiman, B.; Gomez, J.C. Text Matching to Measure Patent Similarity. *Strateg. Manag. J.* **2018**, *39*, 62–84. [\[CrossRef\]](#)
29. Su, H.; Lee, P.-C. Mapping Knowledge Structure by Keyword Co-Occurrence: A First Look at Journal Papers in Technology Foresight. *Scientometrics* **2010**, *85*, 65–79. [\[CrossRef\]](#)
30. Radhakrishnan, S.; Erbis, S.; Isaacs, J.A.; Kamarthi, S. Novel Keyword Co-Occurrence Network-Based Methods to Foster Systematic Reviews of Scientific Literature. *PLoS ONE* **2017**, *12*, e0172778. [\[PubMed\]](#)
31. Xu, H.; Luo, R.; Winnink, J.; Wang, C.; Elahi, E. A Methodology for Identifying Breakthrough Topics Using Structural Entropy. *Inf. Process. Manag.* **2022**, *59*, 102862. [\[CrossRef\]](#)
32. Lim, H.; Kim, Y.; Chun, G.H.; Sohn, S.Y. Identification of Repurposed Smart Healthcare Innovation Opportunities for Cross-Disease Applications via Dynamic Multiplex Disease Network. *Technol. Forecast. Soc. Change* **2026**, *226*, 124573. [\[CrossRef\]](#)
33. Zhai, D.; Zhao, K.; Wang, M.; Zhai, L.; Xu, S. AI-Driven Opportunity Forecasting for Technology Startup Identification: Integrating Graph Embedding, LLMs, and Informetric Analysis. *Technol. Forecast. Soc. Change* **2026**, *227*, 124649. [\[CrossRef\]](#)
34. Hearst, M. Untangling Text Data Mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, MD, USA, 20–26 June 2002. [\[CrossRef\]](#)
35. Feldman, R.; James, S. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*; Cambridge University Press: Cambridge, UK, 2007.
36. He, J.; Nguyen, D.Q.; Akhondi, S.A.; Druckenbrodt, C.; Thorne, C.; Hoessel, R.; Afzal, Z.; Zhai, Z.; Fang, B.; Yoshikawa, H.; et al. ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents. *Front. Res. Metr. Anal.* **2021**, *6*, 654438. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Sarica, S.; Luo, J. Stopwords in Technical Language Processing. *PLoS ONE* **2021**, *16*, e0254937. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Yao, H.; Liu, C.; Zhang, P.; Wang, L. A Feature Selection Method Based on Synonym Merging in Text Classification System. *EURASIP J. Wirel. Commun. Netw.* **2017**, *2017*, 166. [\[CrossRef\]](#)
39. Yang, A.J. Text vs. Citations: A Comparative Analysis of Breakthrough and Disruption Metrics in Patent Innovation. *Res. Policy* **2025**, *54*, 105295. [\[CrossRef\]](#)
40. Aceves, P.; Evans, J. Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory Within Organization Science. *Organ. Sci.* **2023**, *35*, 769–1202. [\[CrossRef\]](#)
41. Rejeb, A.; Simske, S.; Rejeb, K.; Treiblmaier, H.; Zailani, S. Internet of Things Research in Supply Chain Management and Logistics: A Bibliometric Analysis. *Internet Things* **2020**, *12*, 100318. [\[CrossRef\]](#)
42. Newman, M.E.J. Scientific Collaboration Networks. I. Network Construction and Fundamental Results. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2001**, *64*, 016131. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Serrano, M.; Boguñá, M.; Vespignani, A. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 6483–6488. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Kneeland, M.; Aharonson, B. Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation. *Organ. Sci.* **2020**, *31*, 535–795. [\[CrossRef\]](#)

45. Nardo, M.; Saisana, M.; Saltelli, A.; Tarantola, S.; Hoffman, A.; Giovannini, E. *Handbook on Constructing Composite Indicators and User Guide*; OECD Publishing: Paris, France, 2008; Volume 2005, ISBN 978-92-64-04345-9.
46. Dobbie, M.; Dail, D. Robustness and Sensitivity of Weighting and Aggregation in Constructing Composite Indices. *Ecol. Indic.* **2013**, *29*, 270–277. [[CrossRef](#)]
47. Greco, S.; Ishizaka, A.; Tasiou, M.; Torrisi, G. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Soc. Indic. Res.* **2019**, *141*, 61–94. [[CrossRef](#)]
48. Xue, N. Chinese Word Segmentation as Character Tagging. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2003**, *8*, 29–48.
49. Lin, C.; Zhang, Z.-G.; Yu, C.-P. Measurement and Empirical Research on Low-End and New Market Disruptive Innovation. *J. Interdiscip. Math.* **2015**, *18*, 827–839. [[CrossRef](#)]
50. Dotsika, F.; Watkins, A. Identifying Potentially Disruptive Trends by Means of Keyword Network Analysis. *Technol. Forecast. Soc. Change* **2017**, *119*, 114–127. [[CrossRef](#)]
51. Zhang, J.; Yan, Y.; Guan, J. Recombinant Distance, Network Governance and Recombinant Innovation. *Technol. Forecast. Soc. Change* **2019**, *143*, 260–272. [[CrossRef](#)]
52. Liu, X.; Ji, X.; Ge, S. Does the Complexity and Embeddedness of Knowledge Recombination Contribute to Economic Development?—Observations from Prefecture Cities in China. *Res. Policy* **2024**, *53*, 104930. [[CrossRef](#)]
53. Osiyevskyy, O.; Dewald, J. Explorative Versus Exploitative Business Model Change: The Cognitive Antecedents of Firm-Level Responses to Disruptive Innovation. *Strateg. Entrep. J.* **2015**, *9*, 58–78. [[CrossRef](#)]
54. Qiao, Y.; Wang, S. Firms' Structural Positions in Patent Citation Networks and Innovation Performance: Evidence from a Large-Scale Chinese Dataset. *Systems* **2026**, *14*, 351. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.