information

MDPI

Article

From Word Embeddings to Semantic Projections: Interpretability and Context in Web-Scale Semantic Analysis

Mabel López-Bordao [1], Antonia Ferrer-Sapena [1,*], Pablo Lara-Navarra [2,3], Carlos A. Reyes-Pérez [1] and Claudia Sánchez-Arnau [1]

[1] Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, 46022 València, Spain; mlopez2@doctor.upv.es (M.L.-B.); careyper@doctor.upv.es (C.A.R.-P.); csanarn@upvnet.upv.es (C.S.-A.)
[2] Information and Communication Sciences Department, Universitat Oberta de Catalunya (UOC), Rambla del Poblenou 156, 08018 Barcelona, Spain; plara@uoc.com
[3] Open Future Lab, Universidad Tecnológica ECOTEC, Km. 13.5, Samborondón EC092302, Ecuador
* Correspondence: anfersa@har.upv.es

Abstract

Recent developments in NLP and web-scale document analysis have increasingly emphasized the importance of interpretability and contextual dependence in semantic representations. Although modern word embeddings achieve remarkable empirical performance, their semantic structure is often difficult to interpret, since meaning is encoded through latent geometric relations in high-dimensional spaces. This paper discusses an alternative conceptual framework based on explicit contextual semantic relations. Building on ideas from distributional semantics, co-occurrence analysis, and fuzzy set theory, the study revisits semantic projections and related count-based representations as interpretable directional semantic structures for semantic analysis in document corpora and web-based information environments. In this setting, several classical association measures, including PMI and related transformations, may be understood as derived from simpler conditional semantic projections. The methodology is illustrated through a comparative analysis of semantic associations related to "ChatGPT" across general web-scale data and specialized scientific repositories. Our results demonstrate that semantic projections effectively capture persistent contextual structures while remaining sensitive to corpus-specific discourse communities. The resulting perspective emphasizes interpretability, asymmetry, contextual dependence, and direct empirical meaning as central principles for semantic representation.

Keywords: semantic projection; interpretability; natural language processing; fuzzy sets; web-scale analysis

Check for updates

Academic Editors: Damianos P. Sakas and Nikolaos T. Giannakopoulos

Received: 1 July 2026
Revised: 12 August 2026
Accepted: 13 August 2026
Published: 17 August 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) license.

1. Introduction

The mathematical mechanisms used to automatically estimate semantic similarity are embedded, in one form or another, in essentially all NLP (Natural Language Processing) architectures. These systems rely on a broad spectrum of methodologies and theoretical perspectives. Beyond the continuous search for more efficient algorithms, two major concerns have increasingly guided recent research. The first is understanding more precisely how the notion of semantic context intervenes in similarity computations. The second involves improving the interpretability of the produced semantic representations (see, for example, recent contributions on interpretable count-based representations and factorizations of word–context statistics [1]).

This situation suggests a broader perspective. Many apparently different semantic methodologies are, at least partially, transformations of simpler conditional or co-occurrence structures. Thus, it becomes natural to ask whether the true semantic information lies mainly in the final latent representations produced by these methods, or rather in the original contextual relations from which those representations are constructed. From this viewpoint, geometric embeddings may be regarded as transformed versions of simpler and more explicit semantic relations.

The aim of this paper is to discuss this viewpoint from a conceptual and mathematical perspective. Crucially, the paper presents a conceptual and methodological framework supported by our applications, rather than a predictive machine learning model intended to compete on standard NLP benchmarks. Semantic projections are conceived not as a universal replacement for latent embeddings, but as a transparent and auditable complement when contextual specificity and explainability are paramount. The proposed approach builds on classical ideas from distributional semantics and co-occurrence analysis, but also on the original philosophy underlying Zadeh's theory of fuzzy sets and approximate reasoning [2–4]. In Zadeh's formulation, semantic categories are not treated as rigid classes with sharp boundaries, but rather as graded and context-dependent structures whose meaning emerges through partial compatibility relations. This viewpoint anticipated several issues that have become central in current NLP research, including semantic vagueness, contextual meaning, interpretability, and graded semantic representation. Recent surveys on fuzzy methodologies in NLP and explainable AI have in fact highlighted a renewed interest in these ideas [5,6]. Within this general framework, the paper revisits recent developments such as semantic projections and set-word embeddings [7–10] as examples of a more general and interpretable philosophy for semantic representation.

The central idea is simple. Instead of constructing a global embedding space in which semantic meaning is encoded through relative geometric positions, the analysis considers directional semantic relations themselves as the primary objects of analysis. Thus, given a target term v and a finite semantic universe

$$U = \{u_1, \ldots, u_n\},$$

there is associated with v the family of directional semantic indices

$$\mu_v(u_i), \qquad i = 1, \ldots, n,$$

where each quantity measures the degree to which the semantic context of u_i is shared by v. Operationally, these indices are computed through conditional co-occurrence ratios relative to a specified collection of documents or contexts.

The interpretability of such representations is one of the principal motivations of this work. At present, explainability has become a central requirement in the analysis of AI-generated outputs and semantic models. Transparent semantic structures allow researchers to validate conclusions, identify biases and limitations, and transform computational outputs into reproducible scientific knowledge consistent with the scientific method. In contrast, opaque latent representations often hinder the type of logical and methodological scrutiny required for cumulative scientific progress [11].

An important consequence of this philosophy is that many semantic tasks do not necessarily require the construction of a global embedding space. In numerous situations, understanding the relation between a fixed term and a finite semantic universe only requires the direct computation of pairwise semantic sharing indices. In this sense, semantic projections may be viewed as contextual semantic coordinates rather than latent geometric embeddings.

https://doi.org/10.3390/info17080787

The ideas developed in this paper are mainly conceptual and methodological. This work does not propose here a new large-scale embedding architecture, but rather a mathematical and epistemological perspective on semantic representation grounded on interpretability, contextual dependence, and explicit semantic relations. Several recent developments in semantic projections, set-word embeddings, semantic stability analysis, and applications to semantic forecasting already provide concrete realizations of this general viewpoint [7–9,12]. In particular, several practical use cases have already been published as part of the DeflyCompass methodology [12], and preliminary comparisons between semantic projections and other distributional representations have been discussed in previous iterations of this work [9]. Building on these foundations, the present paper focuses on the underlying mathematical and conceptual principles of the framework. Bayesian methodologies may also offer a natural probabilistic framework for incorporating uncertainty and variability into these constructions [13,14], although this aspect will only be briefly discussed here.

2. Related Work

Since the first developments in distributional semantics, the main idea has remained essentially the same: semantic similarity between words is inferred from the contexts they share. In practice, this is usually done through co-occurrence counts and transformations derived from them. Classical collocation measures such as PMI (Pointwise Mutual Information), PPMI (Positive Pointwise Mutual Information), Dice, log-likelihood ratios, t-score, or z-score quantify different types of statistical association between terms and are now standard tools in NLP methodologies (see, for example, [15] and the references therein). Even many modern embedding models that seem formally different still rely on the same underlying co-occurrence information. For example, SGNS (Skip-Gram with Negative Sampling), one of the standard procedures used to construct word embeddings, can essentially be interpreted as a factorization of a shifted PMI matrix [16–18].

These methodologies are now widely used in web-scale semantic analysis, information retrieval, semantic indexing, recommendation systems, and document mining applications. In current web environments, semantic representations are often constructed from extremely large and dynamically evolving textual corpora, making contextual adaptability and interpretability increasingly important requirements [19,20]. More generally, the development of semantic representations for large information systems is closely related to the broader perspective of the Semantic Web and machine-readable semantic structures introduced in [21], highlighting the continuous tension between expressive power and explainability in modern NLP.

Furthermore, recent research has introduced several alternative approaches to interpretability in NLP, such as probing techniques designed to uncover linguistic properties in hidden representations, concept-based explanation models, and sparse embedding representations that restrict activation to interpretable dimensions. Probing methods evaluate whether specific syntactic or semantic attributes can be extracted from latent layers using linear classifiers, while sparse embeddings attempt to enforce disentangled, human-readable semantic axes. While these methods offer valuable insights into the internal mechanisms of high-dimensional models, they often remain disconnected from the direct empirical co-occurrence structures that underpin distributional meaning. In contrast, the semantic projection framework focuses on maintaining an explicit link between the representation and its observable contextual environment, providing a more transparent and auditable basis for semantic analysis.

The proposed viewpoint is also consistent with several recent applications in domain-specific environments where explicit count-based semantic representations remain prefer-

able because of interpretability and governance requirements. Likewise, directional medical embedding models based on asymmetric co-occurrence statistics have shown the importance of conditional and directional semantic relations in real-world contexts [22].

3. Semantic Projections and Conditional Semantic Structures

From a mathematical point of view, the essential object underlying the present approach is a directional semantic relation represented by a conditional quantity. If T is a set of semantic items, the fundamental representation tool is given by the embedding

$$i : T \hookrightarrow \Sigma, \quad i(v) = B_v \in \Sigma,$$

where B_v is, for example, the subset of documents in a document corpus whose content is related in a well-defined way to the semantic item v. As discussed below, this procedure constitutes, broadly understood, the foundational mechanism underlying the definition of semantic projections throughout this work. The underlying idea is straightforward: rather than relying on manually curated definitions or expert annotation, semantic orientation is derived by contextual co-occurrence from the statistical structure of large textual corpora, allowing the resulting representations to reflect genuine patterns of usage. This approach is consistent with the distributional hypothesis and aligns with standard practice in corpus-based and data-driven semantic modeling.

In what follows, two distinct retrieval settings are considered that differ both in scope and in the nature of the corpus they implicitly define. The first setting involves targeted queries submitted to specialized scientific repositories, such as DOAJ or arXiv, which index the peer-reviewed and preprint scholarly literature. In this case, the retrieved documents share a relatively homogeneous register and domain, which tends to produce more focused and technically coherent semantic representations. The second setting involves open-ended queries conducted through general-purpose search engines such as Google, where the retrieved content spans a considerably broader and more heterogeneous range of sources, registers, and genres. While this introduces greater variability, it also captures a wider spectrum of contextual usage, which may in certain cases be desirable. The relative merits of each retrieval strategy will be discussed in relation to the specific tasks and evaluation criteria considered in the applied context explained in Section 6.

The construction and selection of the semantic universe U constitute a critical phase of the methodological pipeline. Typically, U is composed of a finite set of terms that define a thematic domain or represent specific semantic attributes of interest. The choice of these terms depends on the research objectives and the nature of the target concepts being analyzed. In practice, the universe can be validated by evaluating the stability of the resulting projections across different sub-corpora or by ensuring that the selected terms provide sufficient discriminative power to differentiate between target terms. While this paper focuses on the conceptual framework, the optimization of U involves identifying a minimal yet representative set of semantic coordinates that capture the relevant contextual dimensions of the domain.

The methodological pipeline for constructing semantic projections follows a systematic three-stage process. First, the target term v and a finite semantic universe U are identified based on the research context. Second, a document corpus is selected, and co-occurrence counts for the pairs (v, u_i) are extracted using targeted queries. Third, the directional semantic indices $\mu_v(u_i)$ are computed as normalized conditional probability estimates and organized into a vector-like representation (set-word embedding). This pipeline ensures that every coordinate in the final representation remains explicitly traceable to identifiable contextual relations in the source data. Further technical details and mathematical deriva-

tions regarding this pipeline and its comparison with other distributional models can be found in previous iterations of this work [8,9].

As explained in the Introduction, given two semantic items represented by two sets A and B in Σ, the associated semantic projection is defined by

$$\mu_A(B) = P(B \mid A).$$

Conceptually, this quantity measures the proportion of contextual occurrences of A that are shared with B. If co-occurrence counts are computed over a collection of textual contexts (abstracts, papers, and web documents), then

$$\widehat{\mu}_A(B) = \frac{C(A,B)}{C(A)},$$

where $C(A,B)$ denotes the number of contexts in which both terms appear and $C(A)$ denotes the number of contexts containing A.

Within this framework, many classical association measures admit a natural reinterpretation. Recall that

$$PMI(A,B) = \log \frac{P(A,B)}{P(A)P(B)} = \log P(B \mid A) - \log P(B),$$

and therefore

$$PMI(A,B) = \log \mu_A(B) - \log P(B).$$

Thus, PMI appears as a logarithmic transformation of a directional semantic projection corrected by the marginal probability of B. Similarly, positive PMI (PPMI) may be viewed simply as the truncation of this transformed quantity to its positive part:

$$PPMI(A,B) = \max(PMI(A,B), 0).$$

This interpretation suggests that a substantial part of modern semantic modeling can be understood as the manipulation or transformation of explicit conditional semantic structures. In particular, SGNS-type embeddings [16–18] and CA-based semantic decompositions may all be interpreted as different ways of reorganizing or factorizing explicit semantic sharing structures. Semantic projections are the most elementary indices capturing this notion of semantic sharing. In a sense, these indices are understood as the (stochastically defined) "atoms" of the mathematical treatment of semantic relatedness between terms. They are empirically defined objects whose values are entirely determined by the corpus from which they are estimated; taken together with semantic projections derived from other corpora or other methods, they are best understood as samples of a more general abstract object: the "meaning" of the term with respect to a given universe.

Regarding scalability and computational efficiency, the scale and computational complexity of semantic projection calculations are strictly determined by the scope of the underlying retrieval systems and indexed data sources. Because co-occurrence statistics are extracted directly from pre-indexed search engines and scholarly databases (e.g., Google, DOAJ, OpenAIRE), the heavy computational burden of raw document processing, full-text indexing, and high-dimensional space optimization is offloaded to these specialized platforms. For a target term v and a semantic universe U of size n, constructing the set-word embedding requires $n+1$ retrieval queries, yielding a linear computational complexity of $O(n)$ with respect to the universe size. This property makes the framework inherently lightweight, highly scalable, and computationally efficient for web-scale document analysis.

4. Word Embeddings Versus Semantic Projections

Classical word embeddings and semantic projections rely on substantially different philosophies of semantic representation. In standard embedding methodologies, semantic meaning is encoded through the relative geometric position of vectors in a high-dimensional latent space. Semantic similarity is then inferred from distances, angular relations, or algebraic operations between these vectors. This paradigm has proved extraordinarily successful in many NLP tasks, including translation, recommendation systems, semantic retrieval, and large language modeling [10,20].

However, these representations also present several conceptual limitations. First, the semantic meaning of the latent coordinates is usually unclear. Second, many embedding similarities are fundamentally symmetric, while semantic relations are often directional and context-dependent. Third, the resulting representations strongly depend on large optimization procedures whose semantic behavior is not always transparent.

By contrast, semantic projections do not attempt to reconstruct a universal latent semantic geometry. Instead, they represent semantic information through explicit contextual relations relative to a chosen semantic universe and a specified corpus.

The differences between both approaches may be summarized through three main structural aspects:

(1) Mathematical support structure. The basic mathematical framework underlying classical word embeddings is that of normed or metric vector spaces, typically high-dimensional Euclidean spaces. Semantic information is encoded through geometric relations between vectors, and the resulting semantic structure is fundamentally metric. By contrast, set-word embeddings and semantic projections are naturally associated with measure-theoretic structures. The fundamental object is not a vector space, but a measure space equipped with a σ-algebra of subsets representing semantic attributes together with a measure defined on it. Semantic relations are then described through measurable intersections, conditional structures, and normalized sharing coefficients between semantic contents.

(2) Interpretation of semantic coordinates. In standard word embeddings, the coordinates of the vectors do not usually possess a direct semantic interpretation. Meaning emerges globally through the geometry of the latent space and through statistical regularities learned during optimization. Consequently, the semantic role of each coordinate is typically opaque. In semantic projections, however, the coordinates admit an immediate interpretation. Given a target term v and a semantic universe

$$U = \{u_1, \ldots, u_n\},$$

the semantic representation of v is given by the family

$$\mu_v(u_i), \qquad i = 1, \ldots, n,$$

where each coordinate measures the normalized degree of semantic sharing between the target term and the corresponding semantic concept. Operationally, these coefficients may be estimated through normalized co-occurrence relations, for example, by computing the relative size of the intersection of documents in which the involved terms simultaneously appear. Thus, each coordinate directly quantifies a contextual semantic relation and may be interpreted as a degree of semantic compatibility or semantic membership in the original sense of Zadeh's fuzzy-set philosophy.

(3) Contextualization and interpretability. In classical embeddings, contextualization is produced indirectly through the global interaction between semantic items inside the

training corpus and through the optimization process generating the latent geometry. Consequently, contextual information remains largely implicit inside the final representation. In semantic projections, contextualization is introduced directly through the environment used to compute the projections themselves. The semantic representation therefore depends explicitly on the selected document base, corpus, repository, or information source according to the semantic viewpoint one wishes to analyze. Different corpora generate different semantic structures in a direct and controllable way. Nevertheless, in all cases, the interpretability remains explicit: every coordinate continues to represent a measurable degree of semantic sharing between terms.

From this viewpoint, semantic projections should not necessarily be interpreted as competitors of word embeddings, but rather as complementary semantic structures emphasizing interpretability, contextual dependence, and explicit empirical meaning. In many applications, especially those involving semantic analysis, scientific discourse, policy interpretation, web analysis, or explainable AI systems, these properties may be as important as predictive performance itself.

In comparison with established representation learning paradigms (including classical word embeddings, deep contextual language models, sparse embeddings, and probing baselines), set-word embeddings offer distinct mathematical and operational characteristics. While latent models operate in dense Euclidean or tensor metric spaces with opaque coordinates requiring post-hoc geometric analysis or auxiliary probing classifiers, set-word embeddings are built on a measure-theoretic probability space where every coordinate is a directly auditable conditional probability $P(u_i \mid v) \in [0, 1]$. Furthermore, whereas standard distance metrics force symmetric similarities, set-word projections preserve directional conditional asymmetry ($P(A \mid B) \neq P(B \mid A)$), dynamically adapt to selected corpora without retraining, and require $O(n)$ API retrieval queries while natively supporting Binomial confidence intervals and Z-test statistics.

5. Interpretability, Context, and Semantic Representation

One of the principal difficulties of current embedding methodologies is the tension between expressive power and interpretability. High-dimensional embeddings often capture subtle semantic regularities, but the semantic meaning of the resulting coordinates usually remains inaccessible. This opacity is not merely a practical inconvenience: it constitutes a fundamental epistemological limitation when semantic representations are intended to support scientific reasoning, policy analysis, or explainable AI systems, where the validity of conclusions must be traceable to identifiable and verifiable sources of evidence.

Several recent works have highlighted conceptual limitations of standard embeddings. In particular, the use of global geometric relations to encode semantic information may produce unstable or difficult-to-interpret effects [23]. Likewise, recent Bayesian analyses of embedding bias suggest that many standard semantic comparisons involve substantial uncertainty and should not be reduced to single deterministic scores [13]. These observations reinforce the need for semantic frameworks in which uncertainty and variability can be explicitly acknowledged and communicated, rather than absorbed into opaque latent coordinates.

The framework discussed in this paper follows a different philosophy. Semantic meaning is not assumed to emerge from a hidden geometry, but from explicit contextual relations. The semantic universe itself is chosen intentionally, and the representation of a term is always relative to this chosen universe and to the underlying collection of contexts. From a mathematical point of view, this means that the semantic representation of a term is not a fixed vector in an abstract space, but a family of conditional indices whose values depend explicitly on the selected document base and semantic universe. This dependence is

not a weakness of the framework but one of its principal strengths: it makes the contextual assumptions underlying any semantic comparison fully explicit and open to scrutiny.

An important mathematical consequence of this construction is that the resulting representation space inherits a natural probabilistic structure. Each coordinate of the semantic representation is a conditional probability estimate, and therefore lies in the unit interval, admits a direct frequency interpretation, and can be equipped with standard statistical uncertainty quantification. In particular, confidence intervals for individual projection values can be derived from corpus sampling theory, and hypothesis tests can be formulated to assess whether observed differences between semantic profiles are statistically significant or merely reflect corpus variability. This stands in sharp contrast to standard embedding coordinates, whose numerical values do not carry an intrinsic probabilistic meaning and whose uncertainty is rarely quantified in practice.

Consequently, semantic context becomes a controllable parameter rather than an implicit emergent property of a latent space. This makes it possible to compare semantic structures across corpora, temporal periods, scientific domains, or information repositories in a direct and auditable manner. Furthermore, since projection values are computed from observable co-occurrence frequencies, any change in the semantic profile of a term can be traced back to identifiable and interpretable shifts in its contextual associations within the underlying document collection.

Moreover, the resulting representations are naturally asymmetric. This feature is important because many semantic relations are themselves directional. The degree to which the semantic context of "metal" explains the term "gold" is not necessarily identical to the degree to which "gold" explains "metal", since both conditional quantities are normalized by different marginal frequencies. In other words, the directional semantic structure is characterized by the following potential inequality:

$$\mu_A(B) \neq \mu_B(A). \tag{1}$$

This asymmetry is not an artifact of the construction but a faithful reflection of the genuine directional structure of semantic inclusion and specificity. Standard symmetric similarity measures, such as cosine similarity between embedding vectors, necessarily collapse this distinction, treating semantic relations as undirected and therefore losing information that may be relevant for tasks such as hypernymy detection, semantic entailment, or domain-specific terminology analysis. Semantic projections encode this distinction naturally through conditional relations, and the resulting asymmetric structure is both mathematically well-defined and empirically meaningful.

This general philosophy also naturally connects to classical ideas from fuzzy set theory and graded semantic membership [2–4]. In Zadeh's formulation, semantic categories are not regarded as rigid classes with sharp boundaries, but rather as graded and context-dependent structures whose meaning emerges through partial compatibility relations. From a mathematical standpoint, the membership function of a fuzzy set assigns to each element of a universe a value in the unit interval representing its degree of belonging to the category in question. The semantic projection values constructed in this paper have precisely this structure: each projection coefficient assigns to a target term a degree of contextual compatibility with a given semantic concept, normalized to lie in $[0, 1]$ and estimated directly from empirical data. The resulting representation may therefore be understood as a data-driven fuzzy membership function over the chosen semantic universe, one that is neither manually specified nor optimized through a latent objective, but derived transparently from observable co-occurrence statistics.

Interestingly, several recent works in semantic modeling and compositional distributional semantics have also revisited fuzzy and graded semantic structures in modern NLP

settings [5]. Although the present paper does not aim to develop a fuzzy-logic framework in the classical sense, it advocates a related philosophical position: semantic representation should remain explicitly connected to contextual meaning, graded semantic relations, and interpretable empirical structures. In this sense, the approach developed here may be understood as a mathematically transparent realization of the broader epistemological program implicit in distributional semantics: that the meaning of a term is determined by the company it keeps, and that this determination should be made explicit, measurable, and open to critical examination.

6. An Applied Example: Google-Based Semantic Projections and Language Analysis

This section presents the application of the proposed methodology to a concrete case, illustrating how semantic items can be represented within a reduced universe to yield compact, locally defined vector representations in the form of set-word embeddings. A central feature of this approach is that the coefficients comprising these vectors admit a twofold interpretation: on the one hand, they can be regarded as the outcome of a sampling procedure over independent stochastic variables following unknown distributions; on the other, they may be understood as fuzzy membership degrees with respect to a given semantic content, conceived as a fuzzy set. Both readings are fully interpretable and mutually complementary, providing the framework with a principled theoretical grounding from two distinct but compatible perspectives.

Consider, for example, the semantic universe

$$U = \Big\{\text{“trustworthy AI”}, \text{“automation”}, \text{“creativity”},$$

$$\text{“surveillance”}, \text{“productivity”}, \text{“ethics”}\Big\},$$

and let the target term be “ChatGPT”. The selection of these six terms (U) is intentionally designed to cover key dimensions of public, ethical, socio-economic, and technical discourse surrounding Generative AI and Large Language Models: (i) ethical governance (“ethics”), (ii) institutional and regulatory standards (“trustworthy AI”), (iii) technological impacts on labor (“automation”, “productivity”), (iv) human cognitive/artistic capabilities (“creativity”), and (v) societal risks and data governance (“surveillance”). Together, these terms form a balanced, multi-dimensional semantic coordinate system to evaluate contextual shifts across different document environments.

Methodologically, collecting search engine hit counts introduces specific reproducibility challenges that must be explicitly acknowledged. Commercial search engines dynamically update their indices, alter ranking algorithms, personalize outputs based on geographic location and language settings, and often return rounded or estimated result totals rather than exact document frequencies. To maximize transparency and reproducibility, all co-occurrence counts were retrieved during a single collection window in May 2026 using exact-phrase queries enclosed in quotation marks (e.g., “ChatGPT” and “ChatGPT” “ethics”) with global non-personalized regional and language settings.

Using a large web corpus indexed through a search engine such as Google, semantic projections may be estimated through conditional co-occurrence ratios of the form

$$\mu_{\text{ChatGPT}}(u_i) = P(u_i \mid \text{ChatGPT}),$$

where the probabilities are empirically approximated from the relative frequency of indexed documents containing both terms.

Table 1 shows approximate Google-based semantic projections associated with the selected semantic universe along with 95% binomial proportion confidence intervals derived from the underlying corpus sampling. Given the inherent instability of search-engine hit counts, which may fluctuate due to algorithmic updates and temporal shifts, these values should be understood as approximate contextual estimates rather than absolute indices. In this example, the projections corresponding to "productivity", "creativity", and "ethics" are comparatively high, reflecting the strong contextual association of these concepts with current web discourse surrounding ChatGPT. By contrast, the projection associated with "surveillance" is substantially weaker and appears mainly in discussions related to privacy, regulation, or large-scale data collection. The term "trustworthy AI" also presents a relatively small projection value, suggesting that although the concept is important in AI policy and regulation, it still appears less frequently in direct association with ChatGPT within general web discourse.

Table 1. Approximate Google-based semantic projections for the target term "ChatGPT" with 95% binomial proportion confidence intervals. The projections are estimated through normalized web co-occurrence frequencies $\mu_{\text{ChatGPT}}(u) = \frac{\#(\text{ChatGPT} \wedge u)}{\#(\text{ChatGPT})}$, using Google indexed document counts. Here, $\#(\text{ChatGPT}) = 622{,}000{,}000$.

Semantic Term u	**Google(u)**	**Google$(\text{ChatGPT} \wedge u)$**	**$\mu_{\text{ChatGPT}}(u)$**	**95% Binomial CI**
ethics	563,000,000	28,800,000	0.0463	$[0.04629, 0.04632]$
trustworthy AI	76,400,000	823,000	0.0013	$[0.00132, 0.00133]$
automation	767,000,000	15,200,000	0.0244	$[0.02443, 0.02445]$
creativity	905,000,000	30,000,000	0.0482	$[0.04822, 0.04825]$
surveillance	600,000,000	7,710,000	0.0124	$[0.01239, 0.01240]$
productivity	777,000,000	35,200,000	0.0566	$[0.05657, 0.05661]$

An important aspect of this approach is that the resulting semantic structure is inherently directional and contextual. For example,

$$\mu_{\text{ChatGPT}}(\text{ethics}) \neq \mu_{\text{ethics}}(\text{ChatGPT}),$$

since the proportion of documents discussing ethics in the context of ChatGPT differs from the proportion of ethical discussions specifically involving ChatGPT. This asymmetry is difficult to capture through standard symmetric embedding similarities, but appears naturally in the framework of semantic projections.

Moreover, the projections may vary substantially over time, allowing the analysis of semantic evolution in web discourse. Concepts such as "trustworthy AI", "hallucination", or "AI safety" have become increasingly associated with ChatGPT after the public deployment of large language models and the subsequent emergence of debates on factual reliability, regulation, and explainability.

The results obtained from Google, see Figure 1, may also be compared with those produced by specialized scientific repositories such as DOAJ and OpenAIRE, accessed via the DeflyCompass platform. In this case, semantic projections are computed using the peer-reviewed scholarly literature (where the total document count of occurrences for "ChatGPT" in DOAJ is $N_{\text{DOAJ}} = 8998$) instead of the general web indexed by Google. Although the numerical values are not identical, the overall semantic profiles present similar global tendencies across the different corpora. In particular, concepts such as "ethics", "creativity", and "productivity" continue to present comparatively stronger semantic associations with "ChatGPT", while terms such as "surveillance" or "trustworthy AI" remain significantly weaker.

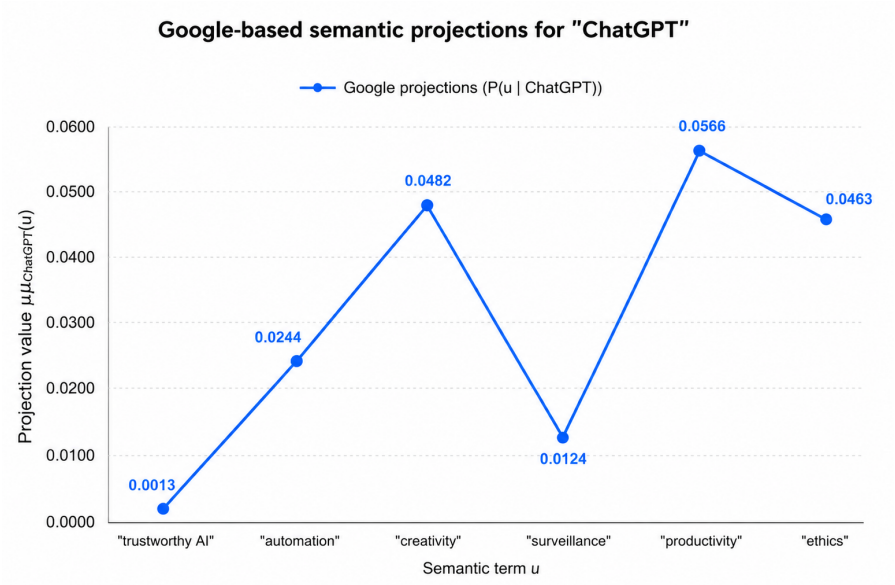


Figure 1. Google-based semantic projections for the target term "ChatGPT" over the selected semantic universe. The values correspond to the conditional semantic projections $\mu_{\text{ChatGPT}}(u) = P(u \mid \text{ChatGPT})$, estimated from normalized web co-occurrence frequencies.

This behavior is natural from the contextual point of view underlying semantic projections. Google indexes a highly heterogeneous web environment containing journalistic articles, blogs, commercial pages, social media discussions, educational materials, and technical documents. Consequently, the resulting semantic projections reflect the broader public discourse surrounding ChatGPT. By contrast, repositories such as DOAJ or OpenAIRE contain mainly scientific and academic publications, where terminology is more specialized, controlled, and contextually constrained. The semantic projections computed from these repositories therefore reflect a more technical and research-oriented discourse.

For instance, the comparatively lower values associated with "trustworthy AI" in the Google corpus may be explained by the fact that this expression belongs mainly to institutional, regulatory, and scientific discussions rather than to general public language. Similarly, concepts such as "surveillance" may appear with different intensities depending on whether the corpus emphasizes public debates, technological policy, privacy issues, or scientific research on AI governance. Nevertheless, despite these contextual differences, the global semantic tendencies remain relatively stable across the different sources. This suggests that semantic projections are capable of capturing persistent contextual semantic structures while still remaining sensitive to the nature of the underlying corpus.

An important consequence of this observation is methodological. Since semantic projections are explicitly conditioned by the selected corpus, they naturally provide a contextual semantic analysis rather than a supposedly universal semantic representation. Different corpora generate different semantic geometries, reflecting the fact that semantic meaning is strongly dependent on the environment in which language is produced. In this sense, semantic projections may be interpreted as contextual semantic observables whose values vary according to the informational ecosystem under consideration.

Moreover, the comparison between Google-based results and scientific repositories illustrates one of the principal advantages of semantic projections with respect to standard latent embeddings: interpretability. Since the projections are directly computed from observable co-occurrence frequencies, variations between corpora can be analyzed and explained in explicit linguistic and contextual terms. The semantic differences are therefore not hidden inside opaque latent coordinates, but remain directly connected to identifiable contextual relations within the corresponding document collections.

Next, Table 2 and Figure 2 allow for an easy understanding of the comparison between different sources for the computation of the semantic projections.

Table 2. Comparison of semantic projection values for "ChatGPT" across Google ($N = 6.22 \times 10^8$), DOAJ ($N = 8998$), and OpenAIRE, including DOAJ 95% confidence intervals and two-proportion Z-test statistics between Google and DOAJ.

Semantic Term	Google Proj. $P(u \mid \text{ChatGPT})$	DOAJ Ratio (Term-Univ. Ratio, 95% CI)	OpenAIRE (Term-Univ. Ratio)	Two-Prop Z-Test (Google vs. DOAJ) (z-Score, p-Value)
"trustworthy AI"	0.0013	0.0050 [0.0035, 0.0065]	0.0005	$z = -9.59$, $p < 0.001$
"automation"	0.0244	0.0190 [0.0162, 0.0218]	0.0090	$z = 3.34$, $p = 0.0008$
"creativity"	0.0482	0.0260 [0.0227, 0.0293]	0.0060	$z = 9.84$, $p < 0.001$
"surveillance"	0.0124	0.0030 [0.0019, 0.0041]	0.0005	$z = 8.06$, $p < 0.001$
"productivity"	0.0566	0.0170 [0.0143, 0.0197]	0.0055	$z = 16.25$, $p < 0.001$
"ethics"	0.0463	0.0475 [0.0431, 0.0519]	0.0160	$z = -0.54$, $p = 0.589$ (n.s.)

The comparison between Google-based semantic projections and DOAJ term-universe ratios reveals both thematic convergences and domain-specific divergences across the terms analyzed. To evaluate whether the observed variations between Google projections ($N_{\text{Google}} = 622 \times 10^6$) and scholarly DOAJ ratios ($N_{\text{DOAJ}} = 8998$) represent statistically meaningful discourse shifts rather than random sampling fluctuations, two-proportion Z-tests were conducted for all universe terms (Table 2). The statistical analysis demonstrates two distinct patterns of contextual association across document environments. On the one hand, cross-corpus semantic convergence is observed for the term "ethics", which exhibits remarkably consistent association across both sources, yielding a projection of 0.0463 on Google and 0.0475 on DOAJ (95% CI: $[0.0431, 0.0519]$). The two-proportion test confirms that this minor difference is not statistically significant ($z = -0.54$, $p = 0.589$), indicating a robust cross-domain consensus: ethical concerns surrounding ChatGPT remain equally prominent whether analyzed in general web discourse or the peer-reviewed scientific literature. On the other hand, all remaining five terms display highly statistically significant differences ($p < 0.05$), capturing structural differences between public and academic discourse priorities. Specifically, terms associated with immediate practical utility, artistic generation, and societal impact are significantly overrepresented in the general web corpus compared to the scientific literature. Productivity shows the largest relative gap (0.0566 vs. 0.0170, $z = 16.25$, $p < 0.001$), followed by creativity (0.0482 vs. 0.0260, $z = 9.84$, $p < 0.001$), surveillance (0.0124 vs. 0.0030, $z = 8.06$, $p < 0.001$), and automation (0.0244 vs. 0.0190, $z = 3.34$, $p = 0.0008$), reflecting the public web's heavy emphasis on workflow optimization, generative capabilities, privacy fears, and labor impacts. Conversely, institutional and regulatory terminology is significantly more concentrated in the peer-reviewed literature: trustworthy AI presents a substantially higher relative projection in DOAJ compared to Google (0.0050 vs. 0.0013, $z = -9.59$, $p < 0.001$), illustrating that formal AI governance, auditing, and safety frameworks constitute a specialized focus of academic research rather than everyday public web conversation.

The values for DOAJ and OpenAIRE are extracted directly from the DeflyCompass platform, ensuring that the reported coefficients represent exact contextual semantic ratios computed from the corresponding repositories. Despite these domain-specific shifts, the overall shape of both curves follows a similar ordinal pattern, which supports the interpretation of semantic projections as probability estimates derived from a sampling process over a given corpus. Figure 2 visualizes these exact semantic projection indices, providing a direct comparison between Google projections and domain-specific DOAJ ratios.

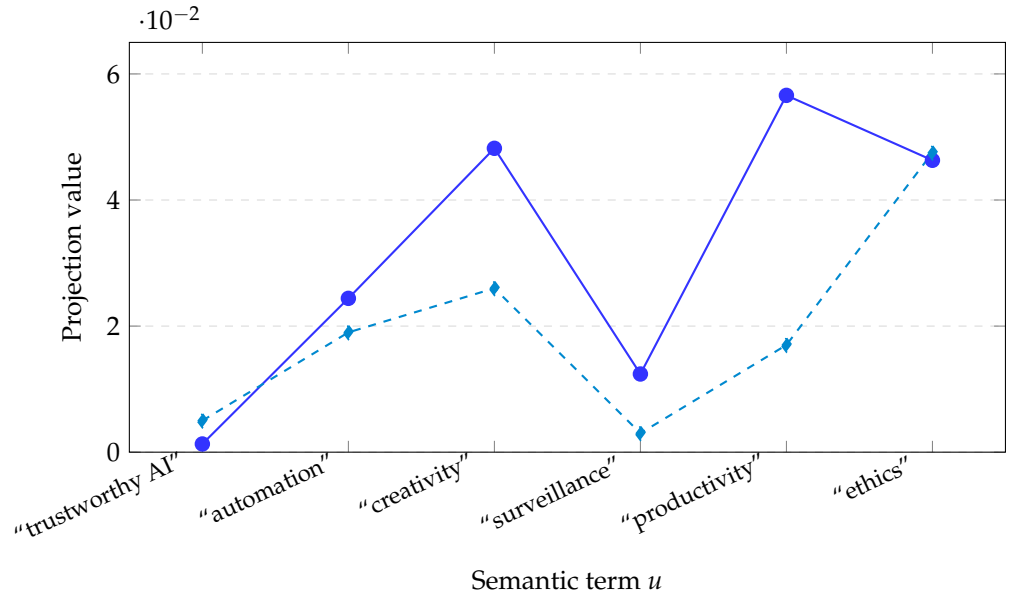


Figure 2. Comparison of semantic projections for "ChatGPT": Google projections are represented by the continuous line, whereas DOAJ projections are shown with a discontinuous line.

A complementary interpretation of the semantic projection values can be drawn from the figures presented above. Rather than viewing the projections as sampling probabilities, they may also be understood as contextual membership coefficients between semantic fields: more precisely, as estimates of the degree to which the semantic field associated with "ChatGPT" belongs to the semantic field of each term in the selected universe. Operationally, these coefficients are computed through normalized co-occurrence ratios relative to the total occurrences of the target term, which introduces an intrinsically directional structure into the resulting semantic relations, and therefore asymmetry. Under this reading, the observed convergences and divergences across sources (e.g., the near-identical values for "ethics" versus the pronounced gap for "productivity") reflect not merely corpus-level frequency differences, but genuine variation in the degree of contextual membership of "ChatGPT" within each semantic field depending on the collection considered.

This interpretation follows the original philosophy underlying Zadeh's theory of fuzzy semantic membership, in which membership degrees are graded rather than binary and depend on the context of evaluation. Modern large-scale document repositories further allow these coefficients to be estimated across markedly different contextual environments. As illustrated in the figures, the projections obtained from Google (reflecting a broad and heterogeneous web discourse) can differ substantially from those derived from scientific repositories via the DeflyCompass platform, where co-occurrence relations are extracted predominantly from academic and technical documents. The contrast observed between both sources, most notably in terms such as "productivity," underscores that semantic projection values are not intrinsic properties of the target concept alone, but are jointly shaped by the nature of the underlying corpus and the semantic context under analysis. This corpus-dependence is itself informative: it reveals how the contextual membership of a concept such as "ChatGPT" shifts across discourse communities, and thus offers a principled means of comparing the semantic positioning of a term in different knowledge domains.

6.1. Word2vec Versus Semantic Projection Measures in an Applied Context

In this section, the mechanism described previously is used to compare the representation using the Word2vec embedding model. The underlying data originates from a previous iteration of this research; readers are encouraged to consult [8,9] for further details.

To evaluate our set-word embedding, see Figure 3, we focus on the eight terms closest to the target term "gold," as ranked by the embedding distance. These terms are: {silver, medal, reserves, coin, rubber, diamonds, tin, timber}. The corresponding distances to "gold" are presented in the second column of Table 3.

In this table, the measures μ are written in "billions" (1,000,000,000). All values in the $\mu(B)$ column have been standardized to this scale to maintain consistent dimensional units. To simplify calculations, we normalize these values by this factor during the distance computation, thereby reducing unnecessary complexity.

In Table 3, term A denotes the target concept ("gold"), while term B represents each candidate word in the universe. The quantities $\mu(A)$ and $\mu(B)$ represent the marginal document occurrence frequencies of A and B. The directional projection $P_A = P(B \mid A) = \mu_A(B)$ measures the conditional probability of encountering concept B in contexts containing A, while $P_B = P(A \mid B) = \mu_B(A)$ denotes the converse conditional probability. The quantities $1 - P_A$ and $1 - P_B$ represent the complementary non-sharing probabilities. Finally, q_μ represents the set-word embedding similarity index based on joint conditional projections and marginal frequencies, as defined in [8,9].

Table 3. Table with the words, the values of the Euclidean distance with the word "gold" and the corresponding values of the elements of the set-word embedding with standardized scales and defined probability symbols.

Word	W2vec	P_A	P_B	$\mu(A)$	$1 - P_A$	$\mu(B)$	$1 - P_B$	q_μ
silver	0.786	0.233	0.476	7.840	0.767	4.110	0.524	8.166
medal	1.129	0.03	0.362	7.840	0.97	0.573	0.638	7.97
reserves	1.135	0.01	0.149	7.840	0.99	0.542	0.851	8.222
coin	1.136	0.046	0.185	7.840	0.954	1.460	0.815	8.670
rubber	1.456	0.027	0.27	7.840	0.973	0.771	0.73	8.191
diamonds	1.109	0.021	0.088	7.840	0.979	0.379	0.912	8.021
tin	1.097	0.032	0.136	7.840	0.968	1.820	0.864	9.161
timber	1.125	0.02	0.21	7.840	0.98	0.288	0.79	8.004

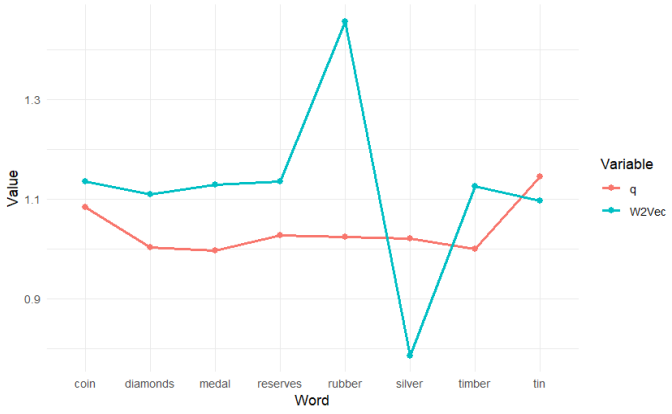


Figure 3. Comparison of the results provided by Word2vec and the set word embedding.

The term "medal" is ranked closer to "gold" than "silver" within the information repository indexed by Google. This outcome is consistent with the high frequency of the bigram "gold medal" in internet corpora, illustrating how common usage patterns dominate over taxonomic similarity. Conversely, while "gold" and "silver" are inherently similar as metals, their linguistic co-occurrence is less frequent than the popularized "gold medal." This suggests that usage, rather than inherent ontological similarity, drives the embedding results. Similar reasoning applies to terms such as "reserves," "rubber," and

"coin." The relationships identified by the set-word embedding for "diamonds," "tin," and "timber" appear more conventional when compared to those generated by Word2vec.

This analysis suggests that when word meanings are derived from their relationships within a semantic environment, they are fundamentally shaped by the measurement tool employed. It appears more practical for many applications to assume that isolated relationships between terms do not fully define a word's meaning; rather, these relationships are inherently contingent upon both context and the specific methodology used to quantify them [8].

Crucially, the differences in these representations highlight the interpretability advantage of semantic projections over classical latent models. While the Euclidean distances produced by Word2vec represent geometric relations in an opaque, high-dimensional space whose individual coordinates lack direct semantic meaning, the coefficients of the set-word embedding (such as $\mu(A)$ and $\mu(B)$ in Table 3) correspond directly to observable contextual relations. Because they are constructed as normalized co-occurrence ratios, these coordinates can be explicitly interpreted as conditional probabilities or fuzzy membership degrees derived directly from the underlying corpus. This transparent mapping from raw data to the final semantic representation ensures that the model is auditable and its contextual dependencies remain explicit, thereby directly addressing the interpretability limitations inherent in standard embeddings.

6.2. Comparative Analysis: Set-Word Embeddings Versus State-of-the-Art Contextual Embedding Models

To further evaluate the empirical behavior of explicit set-word embeddings in comparison with state-of-the-art (SoTA) contextual language models, we examine the target concept "ChatGPT" across the six terms of the semantic universe U (comprising ethics, trustworthy AI, automation, creativity, surveillance, and productivity). We compare the explicit, auditable conditional projections obtained from Google ($N = 6.22 \times 10^8$) and academic DOAJ ($N = 8998$) document corpora against cosine similarity scores derived from six top MTEB benchmark embedding models computed dynamically live via PyTorch/HuggingFace: Sentence-BERT (all-MiniLM-L6-v2), MPNet (all-mpnet-base-v2, BAAI's flagship BGE-Large-v1.5, Mixedbread AI's mxbai-embed-large-v1, Microsoft/Intfloat's e5-large-v2, and BAAI's multi-lingual BGE-M3. To ensure full experimental reproducibility, an open Python 3.11 evaluation pipeline (comparison.py) is provided in Appendix A.

Table 4 presents the comparative empirical results across models and information domains.

Table 4. Empirical comparison of semantic representations for "ChatGPT" across set-word projections (Google and DOAJ) and live top MTEB SoTA contextual transformer models: Sentence-BERT (MiniLM), MPNet-Base, BAAI BGE-Large-v1.5, Mixedbread Mxbai-Embed-Large, Microsoft E5-Large-v2, and BAAI BGE-M3.

Universe Term (u_i)	Google Projection $P(u_i \mid \text{ChatGPT})$	DOAJ Ratio $P(u_i \mid \text{ChatGPT})$	DOAJ 95% CI	Z-Test (p-Val)	MiniLM $\cos(\theta)$	MPNet $\cos(\theta)$	BGE-Large $\cos(\theta)$	Mxbai-Large $\cos(\theta)$	E5-Large $\cos(\theta)$	BGE-M3 $\cos(\theta)$
ethics	0.0463	0.0475	[0.0431, 0.0519]	$z = -0.54$ ($p = 0.589$)	0.118	0.131	0.507	0.483	0.783	0.404
trustworthy AI	0.0013	0.0050	[0.0035, 0.0065]	$z = -9.59$ ($p < 0.001$)	0.070	0.144	0.517	0.481	0.840	0.418
automation	0.0244	0.0190	[0.0162, 0.0218]	$z = 3.34$ ($p = 0.0008$)	0.140	0.165	0.576	0.533	0.824	0.423
creativity	0.0482	0.0260	[0.0227, 0.0293]	$z = 9.84$ ($p < 0.001$)	0.131	0.169	0.545	0.510	0.800	0.385
surveillance	0.0124	0.0030	[0.0019, 0.0041]	$z = 8.06$ ($p < 0.001$)	0.220	0.151	0.556	0.513	0.805	0.403
productivity	0.0566	0.0170	[0.0143, 0.0197]	$z = 16.25$ ($p < 0.001$)	0.170	0.153	0.592	0.546	0.820	0.372

This empirical comparison underscores three essential structural differences between dense vector models and set-word projections.

Comparing dense vector embeddings (Sentence-BERT, MPNet, BGE-Large-v1.5, Mxbai-Embed-Large, E5-Large-v2, and BGE-M3) and set-word projections objectively requires recognizing that their underlying metrics operate in fundamentally different mathematical spaces. Dense models compute unitless cosine similarity scores ($\cos\theta \in [-1, 1]$)

representing spatial proximity in an abstract latent space ($\mathbb{R}^d$), whereas set-word embeddings compute empirical conditional probabilities ($P(u_i \mid v) \in [0, 1]$) measuring relative document co-occurrence. Because direct raw value comparison between a geometric angle and a probability is mathematically invalid, the objective comparison in Table 4 relies on three invariant analytical axes.

First, evaluating ordinal rank order invariance reveals that all six dense contextual models yield top relative similarity alignments towards productivity (≈0.5917 in BGE-Large, 0.5456 in Mxbai-Large, 0.8201 in E5-Large) and automation (≈0.5755, 0.5332, 0.8244), driven by pre-training text distributions and subword lexical alignment. By contrast, set-word projections reveal marked ordinal shifts: public web search (Google) ranks productivity (0.0566) and creativity (0.0482) highest while placing trustworthy AI near the bottom (0.0013), whereas the academic literature (DOAJ) elevates ethics (0.0475) and trustworthy AI (0.0050) while suppressing commercial labor claims (0.0170).

Second, set-word embeddings quantify domain adaptation deltas ($\Delta P = P_{\text{Google}} - P_{\text{DOAJ}}$) across distinct information repositories, demonstrating that public web discourse is over three times more productivity-focused ($z = 16.25$, $p < 0.001$) and nearly four times less focused on trustworthy AI ($z = -9.59$, $p < 0.001$) than the peer-reviewed literature, a dynamic corpus sensitivity that static pre-trained transformer embeddings ($\Delta = 0$) are structurally unable to capture.

Third, while dense neural encoders yield deterministic scalar points without native variance estimates, set-word projections naturally incorporate statistical uncertainty quantification, yielding exact 95% binomial proportion confidence intervals and hypothesis tests to establish rigorous statistical significance.

7. Limitations and Future Work

While the proposed framework emphasizes interpretability and explicit contextual relations, several limitations should be acknowledged. First, semantic projections based on co-occurrence counts are sensitive to corpus bias and the quality of the underlying document collection. In particular, rare-event statistics may be unstable, and the resulting representation depends heavily on the chosen semantic universe. Second, the use of general-purpose search engine counts, while effective for capturing broad public discourse, introduces reproducibility challenges due to the dynamic nature of web indexing and temporal fluctuations. Third, the initial comparative analysis with DOAJ and OpenAIRE relies on examples where values were extracted directly from the DeflyCompass platform. Although these comparisons capture global semantic trends and provide exact numerical indices, they should be interpreted as initial indicators of semantic positioning in discourse communities. Fourth, although the methodology is conceptually scalable to large document collections, its computational efficiency depends on the performance of the underlying retrieval system and the size of the semantic universe.

As highlighted throughout this manuscript, this work presents a conceptual and methodological framework supported by empirical demonstrations and comparative analyses against state-of-the-art dense models (Sentence-BERT, MPNet, BGE-Large-v1.5, Mxbai-Embed-Large, E5-Large-v2, and BGE-M3). A main direction for future empirical research is expanding this comparative framework to large-scale, systematic quantitative evaluation on standardized downstream NLP task benchmarks (such as GLUE, SuperGLUE, and MTEB retrieval suites) to benchmark predictive utility alongside interpretability. Furthermore, systematic quantitative evaluation of the temporal stability and discriminative power of semantic projections across longitudinal web snapshots should be conducted to track the evolution of concepts over time. Finally, future work will explore practical integration

strategies with deep neural architectures, using set-word embeddings as interpretable prior constraints, diagnostic probes, or domain-adapting filters for transformer models.

Because semantic projections are computed as empirical relative frequencies, they naturally support standard statistical uncertainty quantification. Confidence intervals for projection coefficients can be estimated using standard binomial proportion confidence intervals or bootstrap resampling over document corpora, providing an intrinsic structural advantage over opaque latent embeddings.

8. Conclusions

The perspective developed here suggests that interpretable conditional semantic structures may provide a useful alternative or complement to traditional embedding methodologies. Rather than viewing semantic analysis exclusively through the geometry of latent vector spaces, the proposal is to consider explicit contextual relations as primary semantic objects. This reorientation does not merely constitute a technical adjustment, but reflects a deeper philosophical shift: semantic meaning is treated not as a hidden geometric property to be recovered through optimization, but as an explicit, measurable, and context-dependent relation directly grounded in empirical co-occurrence data.

This does not imply that global embeddings should be abandoned. On the contrary, embedding models remain among the most successful tools in NLP and AI, and their predictive performance in tasks such as machine translation, semantic retrieval, and large-scale language modeling is well established. However, many semantic tasks may not require the reconstruction of a universal latent geometry. In numerous contexts, the local and interpretable semantic relations given by the set-word embeddings constructed by semantic projections may already provide sufficient and more transparent information. Furthermore, set-word embeddings offer promising integration strategies with current neural embedding pipelines, where explicit conditional projections can serve as interpretable prior constraints, diagnostic probes, or domain-adapting filters for deep language models. This is particularly relevant in domains where the accountability and auditability of semantic representations are essential requirements.

Web-scale semantic analysis constitutes one of the most natural application domains for this framework. The web is a heterogeneous, dynamic, and contextually stratified information environment in which the meaning of terms shifts continuously across sources, genres, communities, and time periods. Standard embedding models, trained on static corpora and producing fixed latent representations, are inherently limited in their ability to capture this variability in an interpretable and auditable way. Semantic projections, by contrast, are explicitly conditioned on the selected corpus and can therefore be recomputed and compared across different web sources, temporal snapshots, or thematic subcorpora without requiring retraining or architectural modifications. This makes them particularly well suited for tasks such as monitoring the semantic evolution of concepts in online discourse, comparing the contextual positioning of terms across different web communities, or detecting shifts in the associative meaning of a concept following significant real-world events. In information retrieval and semantic indexing applications, the directional and asymmetric structure of semantic projections also offers a natural mechanism for capturing the specificity relations between query terms and indexed content, which symmetric similarity measures are structurally unable to represent.

The applied example discussed in Section 6 illustrates this point concretely. The comparison of Google-based semantic projections with those derived from specialized scientific repositories such as DOAJ and OpenAIRE shows that corpus-dependent semantic structures can be analyzed and interpreted in explicit linguistic terms, without recourse to opaque latent coordinates. The observed convergences and divergences across sources

are themselves semantically informative, revealing how the contextual positioning of a concept shifts across different discourse communities and knowledge domains. This type of analysis would be considerably more difficult to perform and interpret within a standard embedding framework.

Several recent works on semantic projections, set-word embeddings, semantic stability, and contextual semantic analysis provide concrete examples of this direction [7–9]. Bayesian approaches may also offer promising mechanisms for incorporating uncertainty and variability into semantic projections, especially when working with heterogeneous corpora or unstable semantic environments [13,14], although this line of research remains largely open. In particular, the combination of directional semantic projections with probabilistic uncertainty quantification could provide a principled framework for dynamic semantic analysis over evolving corpora.

The framework introduced in this paper advocates a shift from latent geometric semantics toward explicit contextual semantics, emphasizing interpretability, asymmetry, and direct empirical meaning as central principles for semantic modeling. It is suggested that this perspective, far from being in opposition to the dominant paradigm, offers a complementary and necessary viewpoint: one in which the interpretability of semantic representations is not sacrificed in favor of predictive performance, but is recognized as a scientific requirement in its own right.

Author Contributions: Conceptualization, M.L.-B. and A.F.-S.; methodology, P.L.-N. and M.L.-B.; software, C.A.R.-P. and C.S.-A.; validation, C.S.-A.; formal analysis, C.A.R.-P. and P.L.-N.; investigation, M.L.-B. and A.F.-S.; data curation, A.F.-S. and C.S.-A.; writing—original draft preparation, M.L.-B.; writing—review and editing, P.L.-N.; visualization, C.A.R.-P. and C.S.-A.; supervision, A.F.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Generalitat Valenciana (Spain), grant number PROMETEO 2024 CIPROM/2023/32.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

Appendix A. Reproducible Dynamic Evaluation Pipeline Script (comparison.py)

To ensure full experimental reproducibility and facilitate further comparative research, a standalone, open-source Python evaluation pipeline (comparison.py) has been developed. No embedding cosine values are hardcoded or fixed in the script. Instead, the script connects dynamically to local HuggingFace/SentenceTransformers models (e.g., all-MiniLM-L6-v2, BAAI/bge-m3) or local Ollama API instances (such as bge-m3 or nomic-embed-text) to calculate dense vector embeddings and cosine similarities live on the fly during execution.

Appendix A.1. Indications to Run the Pipeline

Running the evaluation pipeline requires Python 3.8+ along with local HuggingFace models (sentence-transformers) or a running Ollama server (https://ollama.ai (accessed on 10 August 2026)). Execution is performed from the terminal command line via `python3 comparison.py -models all-MiniLM-L6-v2 BAAI/bge-m3 -ollama-url` http://localhost:11434 (accessed on 10 August 2026). Upon execution, the script encodes target and universe terms live, computes vector dot-products and L2 norms dynamically,

displays formatted comparison tables in stdout, and exports structured results to evaluation_results.json.

Appendix A.2. Complete Python Source Code

```

#!/usr/bin/env python3
"""
This script dynamically calculates dense vector embeddings and cosine similarities
on the fly using HuggingFace / SentenceTransformers models (e.g. 'BAAI/bge-large-en-v1.5',
'mixedbread-ai/mxbai-embed-large-v1', 'intfloat/e5-large-v2', 'BAAI/bge-m3', etc.)
or local Ollama API instances.

Usage:
python3 comparison.py [--models MODEL1 MODEL2 ...] [--ollama-url URL] [--ollama-model MODEL]
"""

import argparse
import json
import math
import sys
import urllib.request
import urllib.error
from typing import Dict, List, Tuple, Optional, Any

# Ground-Truth Datasets (Contains ONLY empirical co-occurrence data, target terms, and sample sizes)
CHATGPT_DATASET = {
"target": "ChatGPT",
"N_google": 622000000,
"N_doaj": 8998,
"terms": [
{"term": "ethics", "google_projection": 0.0463, "doaj_ratio": 0.0475},
{"term": "trustworthy AI", "google_projection": 0.0013, "doaj_ratio": 0.0050},
{"term": "automation", "google_projection": 0.0244, "doaj_ratio": 0.0190},
{"term": "creativity", "google_projection": 0.0482, "doaj_ratio": 0.0260},
{"term": "surveillance", "google_projection": 0.0124, "doaj_ratio": 0.0030},
{"term": "productivity", "google_projection": 0.0566, "doaj_ratio": 0.0170}
]
}

GOLD_DATASET = {
"target": "gold",
"mu_target_billions": 7.840,
"candidates": [
{"word": "silver",   "w2v_dist": 0.786, "P_A": 0.233, "P_B": 0.476, "mu_B": 4.110, "q_mu": 8.166},
{"word": "medal",    "w2v_dist": 1.129, "P_A": 0.030, "P_B": 0.362, "mu_B": 0.573, "q_mu": 7.970},
{"word": "reserves", "w2v_dist": 1.135, "P_A": 0.010, "P_B": 0.149, "mu_B": 0.542, "q_mu": 8.222},
{"word": "coin",     "w2v_dist": 1.136, "P_A": 0.046, "P_B": 0.185, "mu_B": 1.460, "q_mu": 8.670},
{"word": "rubber",   "w2v_dist": 1.456, "P_A": 0.027, "P_B": 0.270, "mu_B": 0.771, "q_mu": 8.191},
{"word": "diamonds", "w2v_dist": 1.109, "P_A": 0.021, "P_B": 0.088, "mu_B": 0.379, "q_mu": 8.021},
{"word": "tin",      "w2v_dist": 1.097, "P_A": 0.032, "P_B": 0.136, "mu_B": 1.820, "q_mu": 9.161},
{"word": "timber",   "w2v_dist": 1.125, "P_A": 0.020, "P_B": 0.210, "mu_B": 0.288, "q_mu": 8.004}
]
}


# Statistical Utilities
def compute_binomial_ci(p: float, n: int) -> Tuple[float, float]:
if n <= 0: return (0.0, 0.0)
se = math.sqrt((p * (1.0 - p)) / n)
return (max(0.0, p - 1.95996 * se), min(1.0, p + 1.95996 * se))

```

```
def compute_two_prop_ztest(p1: float, n1: int, p2: float, n2: int) -> Tuple[float, float]:
x1, x2 = p1 * n1, p2 * n2
p_pool = (x1 + x2) / (n1 + n2)
se_pool = math.sqrt(p_pool * (1.0 - p_pool) * (1.0 / n1 + 1.0 / n2))
if se_pool == 0: return (0.0, 1.0)
z = (p1 - p2) / se_pool
p_val = 2.0 * (1.0 - 0.5 * (1.0 + math.erf(abs(z) / math.sqrt(2.0))))
return (z, p_val)

def cosine_similarity(vec1: List[float], vec2: List[float]) -> float:
dot = sum(float(a) * float(b) for a, b in zip(vec1, vec2))
n1, n2 = math.sqrt(sum(float(a) * float(a) for a in vec1)), math.sqrt(sum(float(b) * float(b) for b in vec2))
return float(dot / (n1 * n2)) if n1 > 0 and n2 > 0 else 0.0


# Dynamic HuggingFace / SentenceTransformers Evaluator
class DynamicHuggingFaceEvaluator:
def __init__(self, model_name: str):
self.model_name = model_name
self.available = False
self.model = None

try:
from sentence_transformers import SentenceTransformer
print(f"[HuggingFace] Loading live model '{model_name}'...")
self.model = SentenceTransformer(model_name)
self.available = True
print(f"[HuggingFace] Successfully initialized '{model_name}' for live dynamic calculations.")
except Exception as e:
print(f"[HuggingFace Warning] Could not load '{model_name}': {e}")
self.available = False

def get_embedding(self, text: str) -> Optional[List[float]]:
if not self.available or self.model is None:
return None
try:
vec = self.model.encode(text)
return [float(x) for x in vec]
except Exception as e:
print(f"[HuggingFace Error] Failed to encode '{text}': {e}")
return None


# Dynamic Ollama Evaluator
class DynamicOllamaEvaluator:
def __init__(self, base_url: str = "http://localhost:11434", model_name: str = "bge-m3"):
self.base_url = base_url.rstrip("/")
self.model_name = model_name
self.available = self.check_connection()

def check_connection(self) -> bool:
try:
req = urllib.request.Request(f"{self.base_url}/api/tags", headers={"User-Agent": "DynamicEvaluator"})
with urllib.request.urlopen(req, timeout=3) as resp:
data = json.loads(resp.read().decode())
models = [m.get("name") for m in data.get("models", [])]
print(f"[Ollama API] Connected to {self.base_url}. Found models: {models}")
if self.model_name not in models and len(models) > 0:
self.model_name = models[0].split(":")[0]
print(f"[Ollama API] Model selected: {self.model_name}")
return True
except Exception:
print(f"[Ollama API] Server not accessible at {self.base_url}.")
return False

def get_embedding(self, text: str) -> Optional[List[float]]:
if not self.available:
return None
try:
url = f"{self.base_url}/api/embeddings"
```

```
payload = json.dumps({"model": self.model_name, "prompt": text}).encode('utf-8')
req = urllib.request.Request(url, data=payload, headers={"Content-Type": "
    application/json"})
with urllib.request.urlopen(req, timeout=10) as resp:
res = json.loads(resp.read().decode())
return res.get("embedding")
except Exception as e:
print(f"[Ollama Error] Failed to fetch embedding for '{text}': {e}")
return None


def run_dynamic_pipeline(models_hf: List[str], ollama_eval: DynamicOllamaEvaluator)
    -> Dict:
evaluators = []

# Initialize HuggingFace evaluators
for m in models_hf:
hf_eval = DynamicHuggingFaceEvaluator(m)
if hf_eval.available:
short_name = m.split('/')[-1]
evaluators.append((short_name, hf_eval))

# Add Ollama evaluator if available
if ollama_eval.available:
evaluators.append((f"Ollama({ollama_eval.model_name})", ollama_eval))

if not evaluators:
print("\n[ERROR] No live inference engine available!")
print("To run dynamic calculations, either:")
print("  1. Launch Ollama locally (e.g. `ollama pull bge-m3` and `ollama serve`) OR"
    )
print("  2. Install sentence-transformers (`pip install sentence-transformers torch
    `)")
sys.exit(1)

results = {"example_1_chatgpt": [], "example_2_gold": []}

print("\n" + "="*140)
print("LIVE DYNAMIC EVALUATION: 'ChatGPT' Semantic Projections vs. Top MTEB SoTA
    Embedding Models")
print("="*140)

# Encode target 'ChatGPT' across all active evaluators
target_vecs = {}
for name, ev in evaluators:
target_vecs[name] = ev.get_embedding(CHATGPT_DATASET["target"])

for item in CHATGPT_DATASET["terms"]:
term = item["term"]
g_proj = item["google_projection"]
d_ratio = item["doaj_ratio"]
ci_low, ci_high = compute_binomial_ci(d_ratio, CHATGPT_DATASET["N_doaj"])
z_stat, p_val = compute_two_prop_ztest(g_proj, CHATGPT_DATASET["N_google"], d_ratio,
     CHATGPT_DATASET["N_doaj"])

model_cosines = {}
for name, ev in evaluators:
t_vec = ev.get_embedding(term)
if target_vecs[name] and t_vec:
model_cosines[name] = round(cosine_similarity(target_vecs[name], t_vec), 4)
else:
model_cosines[name] = None

entry = {
"term": term,
"google_projection": g_proj,
"doaj_ratio": d_ratio,
"doaj_ci_95": [round(ci_low, 4), round(ci_high, 4)],
"z_stat": round(z_stat, 2),
"p_val": round(p_val, 4) if p_val >= 0.0001 else "<0.001",
"live_model_cosines": model_cosines
}
results["example_1_chatgpt"].append(entry)

# Print Table
```

```
header = f"{'Universe Term':<16} | {'Google P(u|v)':<13} | {'DOAJ Ratio':<10} | {'DOAJ 95% CI':<16} | {'Z-Test (p)':<14}"
for name, _ in evaluators:
header += f" | {name:<18}"
print(header)
print("-" * len(header))

for r in results["example_1_chatgpt"]:
p_str = r['p_val'] if isinstance(r['p_val'], str) else f"p={r['p_val']:.4f}"
ci_str = f"[{r['doaj_ci_95'][0]:.4f}, {r['doaj_ci_95'][1]:.4f}]"
line = f"{r['term']:<16} | {r['google_projection']:<13.4f} | {r['doaj_ratio']:<10.4f} | {ci_str:<16} | z={r['z_stat']:<5.2f} ({p_str})"
for name, _ in evaluators:
val = r["live_model_cosines"].get(name)
val_str = f"{val:.4f}" if val is not None else "N/A"
line += f" | {val_str:<18}"
print(line)

print("="*140 + "\n")
return results


def main():
default_models = [
"sentence-transformers/all-MiniLM-L6-v2",
"sentence-transformers/all-mpnet-base-v2",
"BAAI/bge-large-en-v1.5",
"mixedbread-ai/mxbai-embed-large-v1",
"intfloat/e5-large-v2",
"BAAI/bge-m3"
]
parser = argparse.ArgumentParser(description="Live Dynamic Evaluation Pipeline for Semantic Projections vs. SoTA Embeddings")
parser.add_argument("--models", nargs="+", default=default_models, help="HuggingFace model names to compute live")
parser.add_argument("--ollama-url", default="http://localhost:11434", help="Local Ollama server URL")
parser.add_argument("--ollama-model", default="bge-m3", help="Ollama model name")
parser.add_argument("--output-json", default="evaluation_results.json", help="Output JSON results filename")
args = parser.parse_args()

ollama_eval = DynamicOllamaEvaluator(base_url=args.ollama_url, model_name=args.ollama_model)
results = run_dynamic_pipeline(args.models, ollama_eval)

with open(args.output_json, "w", encoding="utf-8") as f:
json.dump(results, f, indent=2)
print(f"[Export] Dynamic results saved to JSON: {args.output_json}")


if __name__ == "__main__":
main()
```

References

1. Couto, M.; Parapar, J.; Losada, D.E. A study of word embedding models for measuring topic coherence. *Knowl. Inf. Syst.* **2026**, *68*, 160. [CrossRef]
2. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
3. Zadeh, L.A. The Concept of a Linguistic Variable and its Application to Approximate Reasoning. *Inf. Sci.* **1975**, *8*, 199–249. [CrossRef]
4. Zadeh, L.A. From Computing with Numbers to Computing with Words—From Manipulation of Measurements to Manipulation of Perceptions. *IEEE Trans. Circuits Syst. I* **1999**, *45*, 105–119. [CrossRef]
5. Liu, M.; Zhang, H.; Xu, Z.; Ding, K. The Fusion of Fuzzy Theories and Natural Language Processing: A State-of-the-Art Survey. *Appl. Soft Comput.* **2024**, *162*, 111818. [CrossRef]
6. Bordogna, G.; Pasi, G. A Critical Survey on the Use of Fuzzy Sets in Speech and Natural Language Processing. *Int. J. Approx. Reason.* **2014**, *55*, 583–601.
7. Arnau, R.; Coronado Ferrer, A.; González Cortés, Á.; Sánchez Arnau, C.; Sánchez Pérez, E.A. Measuring Semantic Stability: Statistical Estimation of Semantic Projections via Word Embeddings. *Axioms* **2025**, *14*, 389. [CrossRef]

8. Fernández de Córdoba, P.; Reyes Pérez, C.A.; Sánchez Arnau, C.; Sánchez Pérez, E.A. Set–Word Embeddings and Semantic Indices: A New Contextual Model for Empirical Language Analysis. *Computers* **2025**, *14*, 30. [CrossRef]
9. Fernández de Córdoba, P.; Reyes Pérez, C.A.; Sánchez Pérez, E.A. Mathematical Features of Semantic Projections and Word Embeddings for Automatic Linguistic Analysis. *AIMS Math.* **2025**, *10*, 3961–3982. [CrossRef]
10. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2025.
11. Meisenbacher, S.; Chevli, M.; Matthes, F. A Collocation-based Method for Addressing Challenges in Word-level Metric Differential Privacy. In Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, Bangkok, Thailand, 15 August 2024; pp. 39–51.
12. Manetti, A.; Ferrer-Sapena, A.; Sánchez-Pérez, E.A.; Lara-Navarra, P. Design trend forecasting by combining conceptual analysis and semantic projections: New tools for open innovation. *J. Open Innov. Technol. Mark. Complex.* **2021**, *7*, 92. [CrossRef]
13. Dobrzeniecka, A.; Urbaniak, R. A Bayesian Approach to Uncertainty in Word Embedding Bias Estimation. *Comput. Linguist.* **2024**, *50*, 563–617. . [CrossRef]
14. Lauretig, A. Identification, Interpretability, and Bayesian Word Embeddings. In Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, Minneapolis, MN, USA, 6 June 2019; pp. 7–17.
15. McConnell, K.; Blumenthal-Dramé, A. Effects of task and corpus-derived association scores on the online processing of collocations. *Corpus Linguist. Linguist. Theory* **2022**, *18*, 33–76. [CrossRef]
16. Levy, O.; Goldberg, Y. Neural word embedding as implicit matrix factorization. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, 8–13 December 2014; Volume 2, pp. 2177–2185.
17. Melamud, O.; Goldberger, J. Information-Theory Interpretation of the Skip-Gram Negative-Sampling Objective Function. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017*; Barzilay, R., Kan, M.Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2017; Volume 2, pp. 167–171. [CrossRef]
18. Mu, C.; Yang, G.; Zheng, Y. Revisiting Skip-Gram Negative Sampling Model with Rectification. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 1*; Advances in Intelligent Systems and Computing; Arai, K., Bhatia, R., Kapoor, S., Eds.; Springer: Cham, Switzerland, 2019; Volume 997, pp. 485–497. [CrossRef]
19. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
20. Turney, P.D.; Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Intell. Res.* **2010**, *37*, 141–188. [CrossRef]
21. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
22. Wen, J.; Xue, H.; Rush, E.; Panickan, V.A.; Cai, T.; Zhou, D.; Ho, Y.L.; Costa, L.; Begoli, E.; Hong, C.; et al. DOME: Directional medical embedding vectors from Electronic Health Records. *J. Biomed. Inform.* **2025**, *162*, 104768. . [CrossRef] [PubMed]
23. Rogers, A.; Drozd, A.; Li, B. The (Too Many) Problems of Analogical Reasoning with Word Vectors. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 135–148.

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.