

Topical Presence in AI Systems

Prepared March 30, 2026

Important note. This whitepaper does not claim to reproduce Waikay's proprietary formula. The public materials describe topical presence conceptually as a score built from depth, breadth, and concentration of topic associations. The mathematical framework in this document is an academically grounded operationalization of that concept, not a reverse-engineering of Waikay's internal system.

This document seeks to provide a rigorous whitepaper that explains why Waikay's concept of Topical Presence in AI Systems is plausible, measurable, and useful through the lenses of distributional semantics, entity relatedness, topic coherence, and information theory. [1][2][3][4][5][6][7][8]

Topical Presence in AI Systems

Abstract

Topical presence is the measurable degree to which an AI system associates an entity with a defined set of relevant topics. This paper argues that the concept is scientifically defensible because it aligns with well-established research traditions: distributional semantics explains why co-occurrence and context produce semantic association; embedding-based similarity explains how those associations can be measured; topic modeling and topic coherence explain how topical neighborhoods can be evaluated; and Shannon entropy provides a principled way to measure concentration versus diffusion across topics. Based on those foundations, this paper proposes an operational framework for topical presence composed of depth, breadth, and concentration. The result is not a claim about any single vendor's secret formula, but a generalizable measurement model for AI-era brand visibility. [2][3][4][5][6][7][8]

Why the Concept is Plausible

The idea behind topical presence begins with the distributional hypothesis: linguistic meaning can be inferred from patterns of use across contexts. In plain English, words and entities that repeatedly occur in similar environments become semantically related. Reviews of distributional semantics show that vector representations built from co-occurrence or prediction can recover meaningful semantic similarity without manual annotation. That is precisely the scientific grounding for the intuition that an AI system can 'associate' a brand with a topic. [2][3]

Modern NLP systems extend this logic from isolated words to sentences, passages, and entities. Sentence embedding methods such as SBERT make semantic similarity operational by turning text into vectors that can be compared efficiently with cosine similarity. If a brand's surrounding text is consistently close to a set of target topical representations, then the claim that the brand has stronger topical presence than a competitor is measurable rather than mystical. [4][5]

Knowledge graph research adds a second justification. Entity relatedness is not only about literal string matching; it is about the web of relationships linking entities through types, predicates, contexts, and explanatory evidence. Surveys and relatedness research show that semantic proximity among entities can be modeled through heterogeneous relations and graph structures, which aligns with the idea that a brand may be strongly or weakly connected to a category, use case, need state, or attribute cluster. [5][9]

From Semantic Association to Topical Presence

To make the concept useful, we need to move from informal association to a defined measurement problem. Let E be an entity, such as a brand, and let $T = \{t_1, t_2, \dots, t_n\}$ be a curated universe of relevant topics. The job is to estimate how strongly E is associated with each t_i inside an AI-relevant evidence set: training-visible web text, retrieved documents, citations, structured data, reviews, forums, and other knowledge-bearing sources.

For each topic t_i , we can compute an association score a_i between the entity representation and the topic representation. That score may come from cosine similarity between embeddings, a graph-based relatedness measure, a calibrated classifier probability, or a composite signal. Once the association vector $A(E) = [a_1, a_2, \dots, a_n]$ exists, topical presence becomes a problem of summarizing that vector in a meaningful way. [3][4][5][9]

A Proposed Mathematical Framework

The public Waikay materials name three ingredients: depth, breadth, and concentration. Those terms can be formalized in a clean and interpretable way.

| Component | Proposed Definition | Interpretation |
|---------------|---|---|
| Depth | $D(E) = \text{mean}(a_i)$ across all topics or weighted mean across priority topics | Average association strength |
| Breadth | $B(E) = \text{count}(a_i \geq \tau) / n$ | Share of relevant topics with meaningful coverage |
| Concentration | $C(E) = 1 - H(p) / \log n$, where $p_i = a_i / \text{sum}(a)$ | Degree to which association mass is focused rather than diffuse |

Depth measures how strongly the entity is associated with the topic set on average. Breadth measures coverage: in how many relevant topical areas does the entity clear a minimum relevance threshold τ . Concentration measures shape. If association mass is spread evenly across too many weak topics, presence may be diffused; if it is focused on a coherent subset of strategically relevant topics, presence is more concentrated. Shannon entropy is the natural tool here because it quantifies uncertainty or dispersion in a probability distribution. Lower normalized entropy implies greater concentration. [7][8]

A composite topical presence score can then be defined as $TP(E) = w_D \cdot D(E) + w_B \cdot B(E) + w_C \cdot C(E)$, where the weights sum to 1. The weighting can be tuned to business goals. For example, a category leader may care more about breadth, while a niche B2B player may care more about depth and concentration in a smaller commercial topic set.

This framework has two advantages. First, each component is interpretable. Second, it avoids reducing complex semantic behavior to a single opaque number. Analysts can inspect whether a weak total score comes from shallow associations, narrow coverage, or scattered attention.

Measurement Choices and Methodological Considerations

Topic set design matters. Topic universes should be constructed from commercial intent, customer language, product features, category attributes, and competitor framing rather than from arbitrary keyword lists. Topic quality literature shows that interpretability depends on semantic coherence, so poorly defined or overlapping topic sets will degrade the metric. [6][10]

Association estimation also matters. Embedding similarity is attractive because it is flexible and computationally efficient, but semantic similarity scores must be calibrated carefully. Sentence embedding research shows why models like SBERT outperform naïve BERT usage for similarity search, which makes them good candidates for scoring entity-topic closeness. [4][5]

Threshold choice affects breadth. A threshold τ can be global, percentile-based, or topic-specific. Global thresholds make benchmarking easy; topic-specific thresholds better handle heterogeneous categories.

Concentration needs interpretation. High concentration is not always good. A brand that dominates only one microtopic may have very focused presence but poor breadth. That is why concentration should never be interpreted alone; it should always be read alongside depth and breadth. [7][8]

Why Topical Presence is Strategically Useful

Traditional rankings answer a location question: “where did I appear?”

Topical presence answers a meaning question: “what does the system think I am for?”

In LLM-mediated discovery, that second question matters because generative systems synthesize, compare, recommend, and explain rather than merely rank pages. Research on LLM-based recommendation and user trust shows that these systems are increasingly used as recommendation layers, even if users still calibrate trust imperfectly. [11][12]

That makes topical presence useful in at least four ways.

- First, it exposes missing associations, where a brand should belong but does not.
- Second, it reveals displaced associations, where competitors own topics the brand expected to own.
- Third, it gives content and digital PR teams a clearer target than generic 'visibility': build evidence that strengthens specific entity-topic links.
- Fourth, it offers a better diagnostic layer for AI visibility than raw mention counts alone. [1][2][11]

A Practical Workflow for Implementation

1. Define the entity and the market-relevant topic universe.
2. Build text representations for the entity from authoritative evidence: website copy, reviews, documentation, press coverage, citations, forums, analyst commentary, and other model-visible sources.
3. Compute association scores between the entity representation and each topic representation using a validated semantic similarity or relatedness method. [4][5]
4. Derive depth, breadth, and concentration metrics.
5. Compare the resulting profile across competitors, geographies, or product lines.
6. Use the gaps to guide content creation, corroboration, digital PR, citation building, internal linking, and structured evidence design.
7. Recalculate on a schedule and monitor directional change rather than treating any single run as absolute truth.

Limitations

Topical presence is useful, but it is not a perfect readout of model internals. Public-facing measurements usually operate on observable proxies such as retrieved evidence, surfaced responses, embeddings, or citation patterns. They do not reveal the full state of a proprietary model. Topic definitions can also bias results, and embedding-based similarity can overestimate association when language is generic or boilerplate. Because of that, topical presence should be treated as a decision-support metric, not an oracle. [2][4][6]

It is also important to separate semantic association from endorsement. A model may strongly associate a brand with a controversial topic without recommending it. Measurement systems therefore benefit from pairing topical presence with sentiment, stance, or recommendation diagnostics when those distinctions matter.

Conclusion

The main claim of this paper is modest but important: topical presence is not marketing poetry. It can be grounded in established research on semantic representation, entity relatedness, topic quality, and information theory. A sensible implementation turns an entity-topic association vector into three interpretable dimensions - depth, breadth, and concentration - and optionally a composite score. That makes topical presence a credible measurement framework for brands operating in an era where AI systems increasingly mediate discovery, comparison, and explanation. [2][3][4][5][6][7][8][11]

References

- [1] Waikay launch materials describing AI Topical Presence as a metric built from depth, breadth, and concentration, summarized in Newsworthy.ai and related coverage on March 26, 2026.
- [2] T. Cohen and D. Widdows, 'Empirical Distributional Semantics: Methods and Biomedical Applications,' Journal of Biomedical Informatics, 2009.
- [3] D. Jurafsky and J. H. Martin, 'Vector Semantics and Embeddings,' Speech and Language Processing draft chapter, Stanford University, 2023.
- [4] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,' 2019.
- [5] K. S. Brown et al., 'Investigating the Extent to which Distributional Semantic Models Capture Semantic Similarity,' Cognitive Science, 2023; and related semantic similarity literature.
- [6] H. Rahimi et al., 'Contextualized Topic Coherence Metrics,' Findings of EACL, 2024; plus topic-coherence evaluation literature.
- [7] C. E. Shannon, 'A Mathematical Theory of Communication,' Bell System Technical Journal, 1948.
- [8] H. Lei et al., 'Concentrated Document Topic Model,' 2021, using entropy as a measure of topic concentration.
- [9] Y. Zhao et al., 'Representation Learning for Measuring Entity Relatedness with Heterogeneous Information Networks,' IJCAI, 2015; plus surveys on entity and relationship embeddings.
- [10] C. Meaney et al., 'Quality indices for topic model selection and evaluation,' 2023.
- [11] L. Wu et al., 'A Survey on Large Language Models for Recommendation,' 2023.
- [12] M. J. McGrath et al., 'Users do not trust recommendations from a large language model: a preliminary study of LLM-sourced recommendation trust,' Frontiers in Computer Science, 2024.