AI Safety Connect Platform

Jaime Raldúa Veuthey
jaime.raldua.veuthey@gmail.com
Apart Research (Mentor)

Janeth Valdivia
contact@aismx.org
AI Safety México

Dexter Gómez
contact@aismx.org
AI Safety México

Max Pinelo
contact@aismx.org
AI Safety México

Julius A. Odai
jlsodai@gmail.com
Fluent Delta – AI Safety GH

Ihor Kendiukhov
kenduhov.ig@gmail.com
Fluent Delta – AI Safety GH

Tim Sankara
tim.sankara@gmail.com
ORCG

Jakub K. Nowak
jakub.kamil.nowak@gmail.com
AI Safety PL

Kailer Laino
kailer@utexas.edu
The University of Texas at Austin

Kaelan Yim
kaelan@berkeley.edu
University of California, Berkeley

Supervised Program for Alignment Research — Fall 2025

Abstract

AI Safety Connect addresses a structural coordination gap between academic research and AI safety communities by constructing an integrated platform that systematically maps authors, publications, and thematic areas relevant to AI safety. The system relies on a hierarchical taxonomy of Areas, Fields, and Subfields to impose conceptual structure over a heterogeneous research landscape. A comparative evaluation of major scholarly indexers identified Semantic Scholar as the most stable and semantically informative source for large-scale extraction, enabling the retrieval of 185,715 documents aligned with the taxonomy. These data are processed through a Medallion Architecture deployed on AWS, yielding progressively structured representations that include deduplicated metadata, citation graphs, thematic distributions, and author-level profiles. A semantic layer based on E5-Large embeddings provides a high-dimensional representation of conceptual similarity across papers, complementing structural signals derived from citation networks. The platform exposes these components through REST and semantic-search APIs, supporting relevance-based retrieval and researcher matching. By integrating hierarchical querying, graph-based analysis, and embedding-space retrieval within a single computational framework, AI Safety Connect establishes a scalable approach for characterizing the AI safety ecosystem and identifying potential collaborations between academic and community actors.

Keywords: AI Safety Research Ecosystem Mapping, Taxonomy Design, Medallion Architecture, Semantic Search, Data Pipeline.

1 Introduction

Traditional academia provides invaluable intellectual resources for addressing the technical and conceptual challenges of artificial intelligence security. Its researchers possess rigorous methodologies, sophisticated analytical tools, and a well-established tradition of critical thinking. However, a significant portion of this community remains largely unfamiliar with the emerging risks of advanced AI and the recent literature arguing why security should be a central focus of contemporary research agendas.

In parallel, there are communities that operate explicitly under an AI safety mindset, including LessWrong and various groups specializing in AI Safety, have developed a detailed understanding of the risks and necessary mitigation mechanisms. However, these communities often operate simultaneously with conventional academic circuits. This gap represents a structural coordination failure between two groups whose strengths are essentially complementary: traditional academia contributes specialized knowledge, methodological rigor, and technical depth; while AI Safety communities have made substantial contributions in recent years, and their emphasis on deep-seated risk issues and highly technical orientation provide conceptual clarity and sensitivity to risks. The disconnect between the two hinders knowledge transfer, a comprehensive perspective, and limits the creation of collaborations that could be crucial for the development of strategic research in AI safety.

AI Safety Connect emerged as a response to this gap. Its goal is to build a platform that automatically maps who is working on which topics, identifies points of convergence, and suggests connections that reduce friction for meaningful collaborations. More than just an isolated tool, the project is envisioned as a bridge between academic researchers with critical skills for addressing AI safety and communities highly active in discussing and mitigating risks. A helpful analogy to illustrate this vision is that of a "LinkedIn for AI safety": the challenge is not a lack of talent, but a lack of connectivity between those who could work together.

The theory of change guiding this initiative is simple: more accurate connections lead to more efficient research; more efficient research accelerates progress on key security issues; and such progress strengthens our collective preparedness against increasingly powerful and transformative AI systems.

1.1 AI Safety Connect's objectives as a bridge between academia and community

The goal of AI Safety Connect is to reduce coordination friction between traditional academia and AI Safety communities through a discovery and connection system capable of identifying researchers and work aligned with the field's main agendas.

1.1.1 Specific objectives

To achieve our main objective of suggesting meaningful relationships, whether at the article or author level, and facilitating collaborations that would otherwise be difficult to identify, we propose:

- Establish a domain taxonomy that allows systematic coverage of research areas in AI Safety.

- Select reliable extraction methods for academic sources.
- Implement transformation pipelines following resilient data architecture principles.
- Deploy cloud infrastructure for processing serverless.
- Develop API services for programmatic access to data
- Implement semantic search on content of papers and authors.
- Create the user interface for ecosystem discovery and exploration.

2 Methodology

The methodology of this project integrates three complementary processes: (1) a taxonomy and query design as well as systematic evaluation of data sources to ground the platform in reliable and semantically rich inputs; (2) the construction of a reproducible data-processing pipeline based on progressively structured representations; and (3) the development of computational mechanisms for semantic retrieval and researcher matching.

Each component is designed to transform raw, heterogeneous information into increasingly structured and operationally useful representations, following a principle analogous to hierarchical feature extraction in deep learning systems.

2.1 Taxonomy Design, Query and Data Source Evaluation

The taxonomy designed in this project was inspired by A Shallow Review of Technical AI Safety and the paper Open Problems in Technical AI Governance (1; 2), prioritizing the most relevant contemporary research agendas. Based on these references, the taxonomy was constructed to reflect the current state of the art and the main active lines of research within the field of AI safety.

In this sense, the taxonomy presents a structured and multidisciplinary map of AI Safety research, organized into primary thematic areas, each further decomposed into primary and secondary fields, and subsequently into conceptual subfields. Its objective is to systematically capture the diverse theoretical, technical, and normative approaches that contribute to understanding, evaluating, and mitigating the risks associated with advanced artificial intelligence systems.

Taken together, the taxonomy assumes that AI safety is not a purely technical problem, but rather a complex phenomenon that requires the integration of tools from mathematics, computer science, the social sciences, philosophy, economics, public policy, and organizational studies.

Each area explicitly distinguishes between two types of disciplinary fields:

- **Primary Fields:** disciplines that provide fundamental and direct tools for addressing the safety problem under consideration.
- **Secondary Fields:** complementary disciplines that enrich the analysis, enable empirical validation, or connect technical results with human, institutional, and societal contexts.

2.1.1 Hierarchical structure of the taxonomy

The taxonomy is organized into eleven major research areas, structured across three progressive hierarchical levels.

At the first level, the taxonomy defines the core areas aligned with contemporary AI Safety research agendas:

1. Mechanistic Interpretability
2. Scalable Oversight
3. Adversarial Robustness
4. Agent Foundations
5. Alignment Theory
6. Evaluations of Dangerous Capabilities
7. Value Learning and Alignment
8. Cooperative AI
9. AI Governance and Policy
10. Compute Governance
11. Technical AI Governance

The second level consists of 86 disciplinary fields, classified as primary or secondary depending on their role within each area. These fields represent established analytical frameworks through which core AI safety problems are studied.

The third level, comprising 276 subfields, introduces the highest degree of conceptual granularity. At this level, the taxonomy specifies concrete techniques, methods, procedures, and theoretical frameworks that function as minimal units of analysis.

It is at this third level that substantive connections between seemingly disparate areas become visible: many of the most significant links between AI safety and other disciplines emerge not at the level of broad research areas, but through shared mathematical techniques, computational procedures, or highly specialized theoretical frameworks that cut across multiple domains.

2.1.2 Query Design

The hierarchical Area–Field–Subfield schema was used as the basis for query generation and bibliographic metadata extraction. Using combinations of Area, Area + Field, and Area + Subfield, scientific literature aligned with the project's taxonomy was retrieved.

For each identified article, the following bibliographic metadata were extracted and stored:

- Unique paper identifiers (Semantic Scholar, OpenAlex, DOI, arXiv),
- Article title,
- Authors and institutional affiliations,

- Publication year and venue (conference or journal),
- Abstract,
- Links to the original source and PDF,
- Keywords and semantic fields provided by the indexer,
- Citation count.

These metadata were consolidated into a relational structure that supports querying, deduplication, and incremental updates.

2.1.3 Literature extraction

Scientific literature extraction was conducted by evaluating three academic indexers, Google Scholar (via Scholarly), OpenAlex, and Semantic Scholar, with the objective of selecting the most suitable source for the project's use case. Each indexer was assessed in terms of thematic coverage, availability of key metadata (authors, abstracts, identifiers), stability under query rate limits, and its ability to support hierarchical queries derived from the Area–Field–Subfield schema.

2.2 Data Pipeline Construction and Medallion Architecture

The system was implemented as a hierarchical data-flow architecture, in which each processing stage produces a representation of increasing structure and utility.

- **Bronze Layer:** Documents retrieved through taxonomy-guided queries were stored in AWS S3 in Parquet format, partitioned by extraction date and thematic area. This layer preserves the original structure of the data and functions as the immutable entry point for downstream transformations.
- **Silver Layer:** AWS Glue jobs performed deduplication, metadata normalization, and thematic aggregation.
- **Gold Layer:** A final transformation synthesized author-level representations, aggregating publications, thematic distributions, citation statistics, and co-authorship networks.

2.3 Semantic Representation and Retrieval Mechanisms

Beyond structural metadata, the system incorporates a semantic layer designed to model the conceptual proximity between documents. Textual content from Semantic Scholar and LessWrong was normalized and merged into a unified schema containing identifiers, source, consolidated text, timestamps, and fields for vector representations.

Embeddings were generated using the E5-Large model. The corpus was encoded under the "passage" mode, while user queries were encoded under "query," producing a shared latent space in which cosine similarity measures semantic relevance. This dual-encoder structure parallels the way deep models learn distributed representations where conceptual structure emerges as geometric relationships in high-dimensional space.

A lightweight Semantic Search API performs nearest-neighbor retrieval directly in the embedding space, enabling the platform to surface relevant papers and authors even when explicit keyword overlap is minimal.

All components, extraction workflows, transformation pipelines, semantic models, APIs, and visualization endpoints, were deployed within a serverless architecture. The platform updates on a monthly schedule, ensuring that researcher profiles, graphs, and semantic indices reflect the most recent state of the ecosystem.

This methodology enables the transition from raw, unstructured data to a coherent, multi-layered representation of the AI safety research landscape, supporting discovery, analysis, and connection across communities.

3 Development and Implementation

3.1 The origin of the data

The selection of the data extraction source was carried out through a methodical process that, as with other systems, required balancing criteria of coverage, data quality, and technical feasibility. Initially, Google Scholar was evaluated due to its extensive catalog of academic resources; however, it soon became clear that its lack of an official API, along with restrictions associated with automated access, posed significant limitations for a project that required structured and repeatable queries at scale.

In order to base the decision on empirical evidence rather than assumptions, three of the main scientific literature indexers, Google Scholar, OpenAlex, and Semantic Scholar, were systematically compared. The comparison considered multiple dimensions: the breadth of coverage, measured as the number of works aligned with our taxonomy of interest; the structural quality of the metadata; and, especially critically, the availability of full abstracts, given that the abstracts constitute the fundamental unit for the generation of semantic embeddings. Operational aspects such as the presence of duplicates, speed limits (rate limits), privacy requirements, support for public APIs, and the ability to perform queries guided by a hierarchical taxonomy of 11 areas, 86 fields, and 276 subfields, derived from Shallow Review of Technical AI Safety (2024), Open Problems in Technical AI Governance and the specific project documentation also played a decisive role.

Initial experiments showed that, under local conditions, it was possible to identify more than 50,000 unique articles aligned with the taxonomy through hundreds of iterative queries. In the cloud infrastructure, the system reached an even greater scale: 185,715 raw documents were extracted, without any cleaning or deduplication techniques. These figures illustrate both the magnitude of the academic ecosystem related to AI Safety and the need for a robust pipeline capable of cleaning, integrating, and structuring this information before it can be used for semantic analysis or collaboration network inference.

The qualitative results showed marked differences among the three sources. OpenAlex provided a stable API with generous query limits, suitable for continuous pipelines, but its main drawback lies in the quality of the abstracts: many are automatically generated and do not constitute original abstracts, considerably reducing the corpus's usefulness for deep semantic analysis. Although the coverage is broad, its generalist orientation limits the density of works specializing in AI Safety.

Scholarly, on the other hand, exhibited the paradox of offering potentially the most comprehensive coverage through Google Scholar, but in an operationally unfeasible way. After only ten to twenty requests, the service tends to block access, forcing users to

Table 1: Extractor Comparison Matrix

Feature	OpenAlex	Scholarly	Semantic Scholar
API Type	Official API	Unofficial scraping	Official API
Rate Limit	100,000/day	∼100/hour	5 req/sec (with key)
Batch Size	200 papers/page	1 paper/query	1000 papers/bulk
Delay Between Requests	1-3 seconds	60+ seconds	1-2 seconds
Abstract Quality	Generated	Truncated	Full abstracts
Reliability Score	8/10	2/10	10/10
Data Quality Score	6/10	4/10	9/10
Performance Score	9/10	2/10	8/10
Checkpointing	No	No	Yes
Deduplication	Yes	Yes	Yes

Source: Authors' analysis based on official documentation.

introduce pauses of more than a minute between queries. In scenarios where it is necessary to retrieve tens of thousands of articles, this operational fragility makes it impossible to integrate it into a reproducible pipeline, regardless of the breadth of the underlying literature.

Semantic Scholar emerged as the option that achieved the strongest balance between coverage, metadata integrity, and technical stability. Its official API proved consistent and robust, capable of supporting large-scale extraction without interruption, and its specialization in computer science provides a richer representation of the academic ecosystem relevant to AI Safety. This level of reliability is not incidental: the platform is built on a scalable system that organizes scientific knowledge as a heterogeneous literature graph, whose construction relies on NLP tasks to ensure structured and consistent metadata.

Moreover, Semantic Scholar integrates ScienceParse, a high-precision metadata extraction system (F1 of 97.0 for reference-author extraction and 92.1 for paper-author lists), which reinforces metadata integrity and reduces noise in downstream stages of the pipeline. Complementing this, its scientific concept extraction and linking system combines multiple approaches (statistical, hybrid, and off-the-shelf models) and anchors entity linking in manually curated knowledge bases such as UMLS and DBpedia, ensuring richer and more reliable semantic signals (3).

Within this context, the systematic availability of comprehensive abstracts enabled the construction of comparable distributed representations, which are essential for subsequent embedding-based analyses. Thus, the final selection of Semantic Scholar as the primary source is not driven solely by its technical availability through a well-defined API, but by the combination of sufficient coverage, high-quality metadata, semantic richness, and abstract accessibility required to build comparable distributed representations across documents.

3.2 Architectural design in AWS

The system was built following a data-flow architecture with clearly defined interfaces, in which the stages of ingestion, transformation, storage, and access remain separated to preserve process traceability and stability. As a structural principle, the Medallion Ar-

chitecture pattern was adopted, organizing information progressively as it moves through the Raw, Silver, and Gold layers. In the Raw layer, articles obtained from the Semantic Scholar API are stored in Parquet format within S3, partitioned by extraction date and thematic area; this layer constitutes the system's entry point and is fed by a resilient extractor capable of handling pagination, regulating request rates, and ensuring the ordered persistence of data. Ingestion is orchestrated through a periodic EventBridge trigger that activates a monthly Step Functions workflow responsible for traversing the areas defined in the taxonomy, constructing hierarchical queries, and writing the results into the Raw layer, after which a Glue Crawler updates the catalog and SNS notifies the completion of the process.

On this foundation, information is transformed through a series of sequential AWS Glue jobs. The first process performs deduplication and thematic aggregation, unifying articles and assigning area labels. The second enriches each article by retrieving, once again from Semantic Scholar, its inbound and outbound citation relationships. Based on this information, a third process constructs a citation graph that serves as the fundamental structure for revealing patterns of influence and intellectual proximity. A fourth process synthesizes author profiles, estimates their thematic distribution, counts citations, and generates a coherent representation of each researcher within the ecosystem. The resulting Gold layer is exposed through the Authors API, implemented in Lambda and served via API Gateway, enabling low-latency REST queries over Athena. This API constitutes the module responsible for generating researcher recommendations through a citation-based algorithm that ranks authors according to their structural connectivity and the thematic overlap of their publications; by leveraging citation and co-citation relationships derived from the graph, the system estimates academic proximity and prioritizes individuals with higher structural relevance within specific areas of AI safety.

In parallel, an API is under development to enable semantic search over authors and articles through the incorporation of vector representations obtained with embedding models.

The project also incorporates a Demo API that provides unified access to processed information on authors, articles, and relevant topics within the ecosystem, integrating data from Semantic Scholar and LessWrong to support search, listings, co-authorship analysis, karma metrics, and citation visualizations. The system currently integrates data from Semantic Scholar, with future ingestion of LessWrong content planned under the same taxonomy of areas, fields, and subfields, and exposes its results through a Next.js frontend designed for corpus exploration and network visualization.

Overall, the architecture operates through infrastructure as code in Pulumi, which manages the creation and updating of the system's critical resources, including storage buckets, Glue databases, containerized Lambda functions, Step Functions workflows, EventBridge schedules, API Gateway endpoints, and Athena queries. The use of least-privilege IAM roles, together with the declarative nature of the deployment, ensures system reproducibility and allows its modules to evolve without compromising overall coherence.

Taken together, this architecture transforms an initially scattered collection of approximately 185,715 articles into a coherent representational system: a structured network of authors, topics, and citation relationships that enables the AI safety ecosystem to be observed from a computational, scalable, and collaboration-oriented perspective.

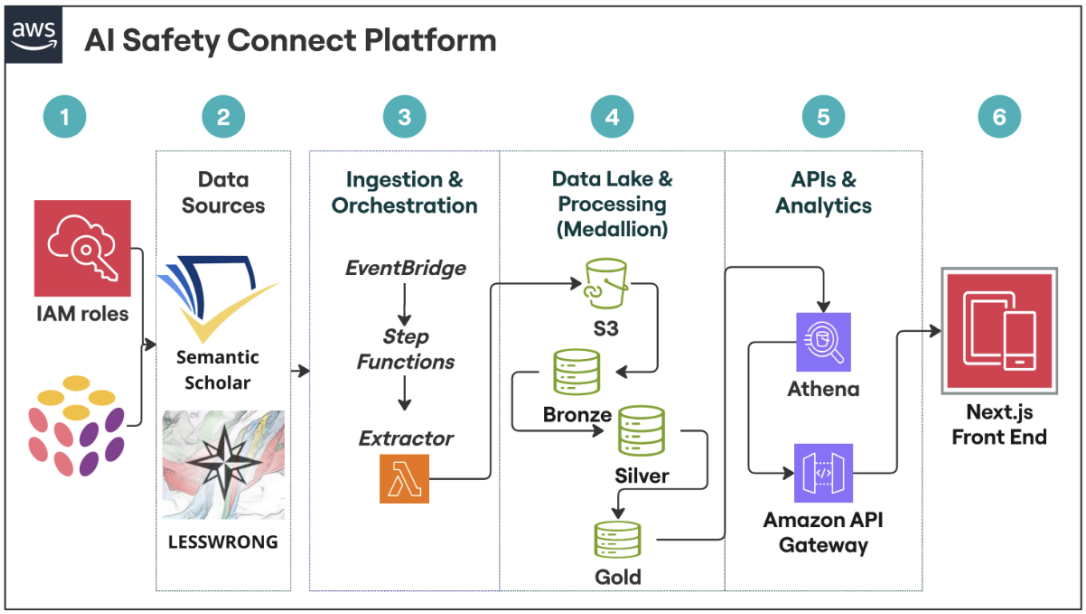

Figure 1: Source: Authors' proposed architecture.

3.3 Semantic Search

To enable semantic search within the AI Safety ecosystem, a system was developed capable of constructing distributed representations of academic and community content, analogous to how language models learn deep structures from textual data. The central idea is to transform each document, whether it's the abstract of a scientific article or a LessWrong publication, into a high-dimensional vector that captures its conceptual meaning, allowing similarity operations that reveal relationships difficult to detect through literal matches.

The process begins with data extraction from Athena, where the different sources are consolidated using specialized scripts. LessWrong publications include metadata on activity, subject tags, and textual content in Markdown and plain text formats, while records from Semantic Scholar contribute abstracts, citation metadata, affiliations, fields of study, and publication dates. These sets, heterogeneous in structure and purpose, are then integrated through a reproducible workflow that unifies both sources and produces a coherent corpus.

As part of the semantic representation process, the next visualization illustrates how the embedding space organizes papers according to their conceptual similarities. Using titles and abstracts, we create vector representation, and the projection provides an intuitive demonstration of how thematic proximity appears within distributed embeddings.

This integrated collection forms the basis for generating semantic representations. Each document is normalized into a common schema that includes its identifier, source, consolidated text, category, presence or absence of a summary, timestamp, and the field designated for storing its embedding. The E5-Large model is used to create these representations. This system is trained to produce comparable vectors using a dual mechanism: the keyword "passage" is used to build the corpus's vector base, while "query" is applied exclusively to the text entered by the user in subsequent searches. This parallel design allows the semantic space to maintain coherence between the corpus and queries, a principle analogous to the concurrent learning of representations observed in deep language

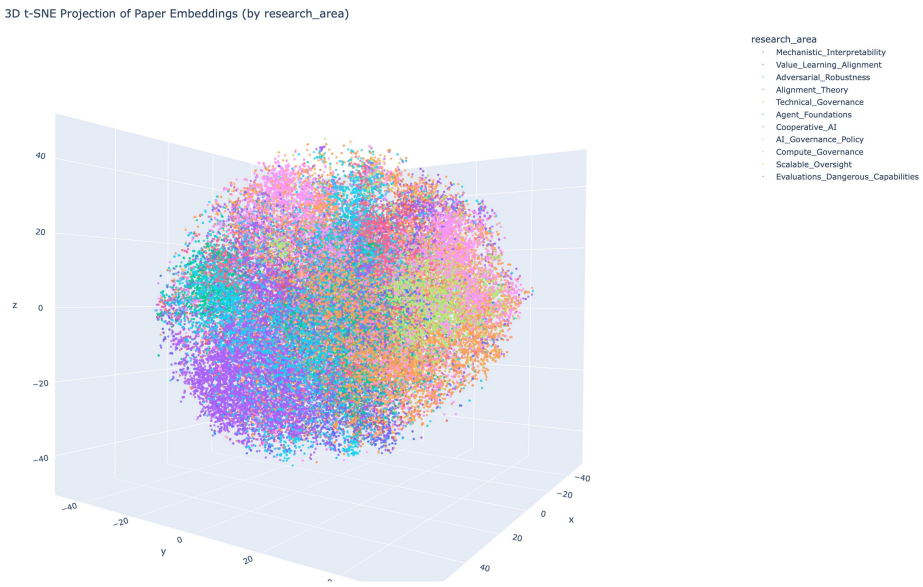

Figure 2: 3D t-SNE Visualization of Paper Embeddings by Research Area. Source: Authors' desing.

understanding networks.

Once the vector space is generated, the system exposes its capabilities through a lightweight API designed to retrieve relevant documents using cosine similarity in the embed space. In this sense, a query like "Technical AI Governance" allows the identification of works that not only share terminology but also conceptual proximity to the topic.

This module operates as the semantic counterpart to the structured pipeline described above: while the Medallion architecture organizes data according to quality levels and explicit transformations, the embeddings layer organizes the corpus's implicit knowledge into a continuous space where thematic and conceptual affinities emerge naturally from the text's statistics. Together, these two approaches allow us to view the AI Safety ecosystem from two complementary perspectives: the explicit structure captured by citations, authors, and areas, and the latent structure revealed by the distributed meaning of the documents.

4 Discussion

AI Safety Connect addresses a highly relevant coordination problem through a scalable and technically viable approach. During the initial phases of the project, scraping techniques were explored as a preliminary strategy for data acquisition, proving effective for exploratory experiments and early validation of the approach. However, when assessed at larger scale, this method revealed significant limitations, particularly with respect to privacy concerns, fragility to changes in data sources, and a lack of resilience for automated and reproducible maintenance.

In response to these limitations, and with the aim of aligning the system with principles of privacy, traceability, and operational sustainability, an approach based on the use of official APIs for data extraction was adopted. This decision enabled the establishment of more stable ingestion workflows that are ethically robust and compatible with a peri-

odic processing architecture, providing a more appropriate foundation for the long-term evolution of the system.

In its current state, the platform operates as a fully functional end-to-end system, with regular extraction processes, structured data-processing pipelines, and access services that support queries at both the article and author levels. These components provide a stable basis for the systematic exploration of the AI Safety research ecosystem and for generating recommendations grounded in structural signals derived from citation and co-authorship networks.

At present, the system is undergoing a phase of progressive integration between its access and visualization layers. While the backend infrastructure has advanced in exposing services that retrieve author profiles and relationships derived from citation networks, the user interface is being adapted to directly consume these services over the deployed infrastructure. This consolidation phase is critical for validating the coherence between the system's structural and semantic representations, as well as for preparing the integration of semantic search mechanisms that enable natural language queries over the integrated corpus.

Beyond being viewed as a technical implementation project, from a research perspective AI Safety Connect can be understood not only as an applied platform, but as an instrument for the empirical study of the structure and dynamics of the AI Safety research ecosystem. The availability of large-scale structural and semantic representations opens the possibility of formulating and evaluating fundamental questions about scientific coordination, such as which types of signals predict meaningful collaboration, how intellectual influence is distributed across subfields, or how the structure of the ecosystem shapes the pace and direction of progress on core safety problems. In this sense, the system constitutes an experimental foundation for analyzing coordination and collaboration as measurable and modelable phenomena.

5 Conclusion

AI Safety Connect addresses a critical coordination gap between traditional academic research and AI safety communities. The platform maps 185,715 documents across a hierarchical taxonomy of 11 areas, 86 fields, and 276 subfields, creating a computational foundation for identifying collaborations and reducing friction in knowledge transfer.

The comparative evaluation of academic indexers established Semantic Scholar as the most suitable source, balancing coverage, metadata integrity, and operational stability. The Medallion Architecture deployed on AWS transforms raw bibliographic data into progressively structured representations, from immutable storage through normalized layers to author-level aggregations. Semantic search through E5-Large embeddings complements citation-based signals, enabling discovery of relevant work even when terminology differs across communities.

We expect the integration of structural and semantic signals to become far more powerful as the system matures. Citation graphs reveal influence patterns, while distributed representations capture conceptual proximity that transcends explicit keyword matches. Together, these mechanisms enable the platform to function both as an applied tool for discovery and as an empirical instrument for studying how scientific communities coordinate around critical problems.

The taxonomy, while comprehensive, must evolve with the research landscape. We

expect structured governance processes to systematically incorporate emerging research areas and subfields, ensuring the system remains responsive to an evolving field. The semantic layer will expand beyond LessWrong to include additional community sources, forums, and gray literature that capture the full breadth of AI safety discourse.

Natural language queries over the integrated corpus are in their early stages, but we expect semantic search to become much more effective as the embedding space incorporates more diverse sources and as query understanding improves. The combination of hierarchical taxonomy navigation and semantic retrieval will enable users to explore the ecosystem from multiple perspectives: following citation networks, traversing thematic areas, or discovering conceptual connections through natural language.

Ultimately, major progress in addressing AI safety coordination challenges will come about through systems that combine structural analysis with semantic understanding. The platform's dual nature, as both a discovery tool and a research instrument, opens possibilities for empirical investigation of fundamental questions: Which signals most accurately predict meaningful collaboration? How does ecosystem structure shape progress on core safety problems?

As AI systems become increasingly powerful and transformative, the need for efficient knowledge transfer between academic and community actors will only intensify. AI Safety Connect represents a concrete step toward building the infrastructure necessary to meet that need.

References

References

[1] "Shallow Review of Technical AI Safety, 2024", LessWrong, December 29, 2024. https://www.lesswrong.com/posts/fAW6RXLKTLHC3WXkS/shallow-review-of-technical-ai-safety-2024 (accessed December 7, 2025).

[2] Reuel, A., et al. (2024b). Open problems in technical AI Governance. [Full citation to be added]

[3] W. Ammar et al., "Construction of the Literature Graph in Semantic Scholar," Allen Institute for Artificial Intelligence, 2018.

[4] Semantic Scholar API Documentation. https://api.semanticscholar.org/

[5] OpenAlex API Documentation. https://docs.openalex.org/

[6] "AI Safety Connect", GitHub. https://github.com/AI-Safety-Connect

A Query Construction Examples

Example queries generated from the taxonomy structure:

- "Mechanistic Interpretability" "Neuroscience"
- "Scalable Oversight" "Game Theory"
- "Alignment Theory" "Multi-objective optimization"

The hierarchical query construction follows this pattern: for each area, generate queries combining the area name with each field name (Area + Field) and each subfield name (Area + Subfield). This approach generates approximately 373 queries.

B Taxonomy Example

Table 2: Taxonomy Structure Example

Area	Field_Type	Field	Subfield
Mechanistic Interpretability	Primary	Neuroscience	Neural coding
Mechanistic Interpretability	Primary	Neuroscience	Connectomics
Mechanistic Interpretability	Primary	Neuroscience	Neural decoding

Table 2 – continued from previous page

Area	Field_Type	Field	Subfield
Mechanistic Interpretability	Primary	Neuroscience	Brain imaging analysis (fMRI/EEG interpretation)
Mechanistic Interpretability	Primary	Signal Processing	Sparse coding
Mechanistic Interpretability	Primary	Signal Processing	Dictionary learning
Mechanistic Interpretability	Primary	Signal Processing	Blind source separation
Mechanistic Interpretability	Primary	Signal Processing	Independent component analysis
Mechanistic Interpretability	Primary	Applied Mathematics	Matrix factorization
Mechanistic Interpretability	Primary	Applied Mathematics	Tensor decomposition
Mechanistic Interpretability	Primary	Applied Mathematics	Spectral analysis
Mechanistic Interpretability	Primary	Applied Mathematics	Wavelet analysis
Mechanistic Interpretability	Primary	Statistical Physics	Phase transitions in neural networks
Mechanistic Interpretability	Primary	Statistical Physics	Mean field theory
Mechanistic Interpretability	Primary	Statistical Physics	Renormalization group methods
Mechanistic Interpretability	Secondary	Computer Vision	Feature visualization
Mechanistic Interpretability	Secondary	Computer Vision	Saliency maps
Mechanistic Interpretability	Secondary	Computer Vision	Attention mechanisms
Mechanistic Interpretability	Secondary	Computational Linguistics	Probing tasks
Mechanistic Interpretability	Secondary	Computational Linguistics	Linguistic structure discovery
Mechanistic Interpretability	Secondary	Computational Linguistics	Syntax trees in transformers
Mechanistic Interpretability	Secondary	Systems Biology	Network motif detection
Mechanistic Interpretability	Secondary	Systems Biology	Pathway analysis
Mechanistic Interpretability	Secondary	Systems Biology	Modular decomposition
Mechanistic Interpretability	Secondary	Information Theory	Mutual information

Table 2 – continued from previous page

Area	Field_Type	Field	Subfield
Mechanistic Interpretability	Secondary	Information Theory	Information bottleneck theory
Mechanistic Interpretability	Secondary	Information Theory	Compression analysis