

Discovering Hierarchical Semantic Structures in EU Regulatory Embeddings: A Comparison of Two K-means Solutions and Their Alignment with EuroVoc

Kayoko Fukasawa
Honda Motor Co., Ltd.
kayoko_fukasawa@jp.honda

Abstract

EU regulations govern interconnected domains, yet the hierarchical organization and granularity of regulatory semantic structures remain insufficiently understood. We analyze 717 provisions from nine EU regulations, represented as 746 embedding units, using K-means and Ward linkage. Two independently optimized 80-cluster solutions reveal recurring domain-specific and cross-regulatory themes but differ in hierarchical organization. Cross-solution analysis shows high centroid-level semantic correspondence (mean matched-centroid cosine similarity = 0.943) despite lower partition agreement (ARI = 0.520; mean membership Jaccard similarity = 0.492). We identify 10 split and nine merge candidates, suggesting potential local differences in semantic boundaries or granularity. Post-hoc EuroVoc alignment shows partial agreement with established categories and identifies two qualitative candidate branches spanning taxonomy boundaries. These findings suggest that broadly corresponding regulatory semantic structures can coexist with local differences in membership, hierarchy, and granularity.

1 Introduction

Environmental sustainability, digital platforms, data governance, AI, and healthcare are governed by interconnected EU legal frameworks. Understanding relationships among regulatory requirements across these domains is challenging because legal provisions encompass domain-specific topics and cross-regulatory functions.

Text embeddings represent legal provisions in a shared semantic space, but their hierarchical organization and granularity across alternative clustering solutions remain insufficiently understood. Similar semantic themes may emerge across solutions despite differences in cluster membership and semantic boundaries.

We investigate these issues using 717 provisions from nine EU regulations, represented as 746 embedding units. We discover hierarchical semantic structures using K-means and Ward linkage and compare two independently optimized 80-cluster solutions. We distinguish partition agreement from centroid-level semantic correspondence and examine potential local differences in semantic boundaries and granularity through centroid similarity distributions.

Finally, we compare the induced hierarchies with EuroVoc as an external reference taxonomy to examine their alignment with established categories and potential cross-taxonomy relations.

1.1 Research questions

We address the following research questions:

- **RQ1: Semantic discovery.** What latent semantic structures emerge from unsupervised embedding-based clustering of heterogeneous EU regulatory documents?
- **RQ2: Cross-solution correspondence and granularity.** To what extent do two independently optimized k-means solutions agree in cluster membership and centroid-level semantic correspondence, and what potential differences in local semantic boundaries and granularity are suggested by their cross-solution correspondence patterns?
- **RQ3: Hierarchical alignment with EuroVoc.** To what extent do the discovered semantic structures align with EuroVoc across concept, Domain, and Microthesaurus levels, and where do they reveal cross-taxonomy relations?

1.2 Contributions

Our contributions are threefold:

1. **Hierarchical semantic discovery across heterogeneous EU regulations.** We identify recurring domain-specific and cross-regulatory semantic themes across two independently optimized K-means hierarchies and characterize differences in their organization.
2. **Exploratory analysis of partition agreement and semantic correspondence.** We distinguish partition agreement (ARI and NMI) from centroid-level semantic correspondence and identify candidate split-merge patterns suggesting potential local differences in semantic boundaries and granularity.
3. **Multi-level alignment with EuroVoc.** We compare the induced hierarchies with EuroVoc at the concept, Domain, and Microthesaurus levels to examine established-category alignment and potential cross-taxonomy semantic relations.

2 Related Work

Domain-specific language models have substantially advanced legal NLP. LEGAL-BERT demonstrated the effectiveness of adapting pretrained language models to legal corpora (Chalkidis et al., 2020). EURLEX57K and MultiEURLEX addressed EU legal document classification using predefined EuroVoc descriptors (Chalkidis et al., 2019, 2021). In contrast, our study investigates semantic structures emerging from regulatory embeddings without using predefined taxonomies during clustering.

Unsupervised topic and concept discovery has increasingly combined contextual embeddings with clustering. BERTopic, for example, uses transformer-based embeddings to identify and describe document clusters (Grootendorst, 2022). Hierarchical topic discovery has also been explored using box embeddings to represent topic taxonomy relationships (Lu et al., 2024). Building on embedding-based semantic discovery, we apply K-means clustering and Ward linkage to discover hierarchical semantic structures from heterogeneous EU regulatory provisions. We use centroid-nearest provisions to support the interpretation of regulatory concepts at multiple levels of abstraction, including regulation-specific and cross-regulatory semantic groups (RQ1).

Clustering stability research has examined the reproducibility of clustering solutions and the sta-

bility of individual clusters (Hennig, 2007; von Luxburg, 2010).

Partition agreement can be assessed using ARI and NMI (Hubert and Arabie, 1985; Vinh et al., 2010), while topic modeling research has investigated correspondence across alternative solutions (Marani and Baumer, 2023). We distinguish partition agreement from centroid-level semantic correspondence and explore potential differences in local semantic boundaries and granularity (RQ2).

EuroVoc-based legal classification relies on predefined descriptors to organize legal documents (Chalkidis et al., 2019, 2021; Avram et al., 2021). In contrast, we introduce EuroVoc only after clustering as an external reference taxonomy. We compare the induced hierarchies with EuroVoc at the concept, Domain, and Microthesaurus levels to examine both alignment with established categories and candidate cross-taxonomy semantic relations (RQ3).

3 Data

3.1 Regulatory Corpus

The final corpus comprises 717 unique, non-empty provisions from nine EU legal instruments: the Artificial Intelligence Act (AI Act), General Data Protection Regulation (GDPR), Data Act, Digital Services Act (DSA), Digital Markets Act (DMA), Ecodesign for Sustainable Products Regulation (ESPR), European Health Data Space Regulation (EHDS), EU Taxonomy Regulation, and Battery Regulation.

The English-language versions of these instruments were collected from EUR-Lex, the official online repository of European Union law. We use the term *provision* to refer to an individual regulatory article used as the unit of analysis. Recitals, annexes, and document-level metadata were not included in the clustering corpus.

The instruments cover heterogeneous but partially overlapping regulatory areas, including AI governance, data protection, data access, digital platforms, health data, sustainable finance, product sustainability, and battery governance. This diversity allows us to examine whether the embedding space captures both regulation-specific topics and cross-regulatory functions.

Table 1 summarizes the composition of the corpus. The dataset contains 249,368 words, with an overall mean of 347.8 words per provision. Corpus size varies across the instruments, from 27 provi-

sions in the EU Taxonomy Regulation to 113 in the AI Act. Mean provision length ranges from 296.8 words in the GDPR to 381.3 words in the Data Act.

3.2 Preprocessing

The extracted texts were normalized by removing duplicated whitespace and non-substantive formatting elements. Empty and duplicate provisions were then removed, yielding the final corpus of 717 unique, non-empty provisions reported in Section 3. Each retained provision was associated with its source instrument, CELEX identifier, article identifier, and original text, thereby preserving document-level provenance throughout the analysis.

Provisions exceeding 1,500 tokens were divided into chunks using the Qwen3-Embedding-8B tokenizer. The chunking procedure prioritized legal structure, including paragraph, numbered-item, and sentence boundaries, with a target length of 1,024 tokens and a maximum of 1,500 tokens.

Of the 717 provisions, 16 were split, yielding 746 embedding units. The resulting units had a mean length of 423.47 tokens, a median of 333.50 tokens, and an interquartile range of 158.25–607.75 tokens. The maximum length was 1,449 tokens.

Each chunk retained its source provision identifier to preserve provenance. Embedding and clustering were conducted on the 746 units.

Word counts were calculated for the 717 retained provisions after text extraction, normalization, and duplicate removal, but before chunking. They are reported only as descriptive corpus statistics and should not be interpreted as model-token counts.

4 Method

Figure 2 in Appendix A provides an overview of the proposed framework, including hierarchical semantic discovery, cross-solution correspondence analysis, and post-hoc EuroVoc alignment.

4.1 Semantic clustering

We represent each of the 746 preprocessed provision units using Qwen3-Embedding-8B, a multilingual text-embedding model from the Qwen3 Embedding series (Zhang et al., 2025).

No task-specific instruction prefix was added; each provision chunk was encoded directly using its text. We used the model’s last-token pooling procedure and applied L2 normalization to the resulting 4,096-dimensional embeddings.

We fixed the number of clusters at $K = 80$ as an exploratory resolution informed by preliminary experiments. With 746 embedding units, this setting yields approximately nine units per cluster on average. Our aim was to obtain fine-grained candidate concepts that could be interpreted through representative provisions and subsequently organized into broader hierarchical semantic structures using Ward linkage. We treat $K = 80$ as a design choice rather than an empirically optimal number of clusters.

The normalized embeddings were clustered using k-means with `n_init=10` and random-state values of 42 and 12. For each random state, the implementation performed 10 internal initializations and retained the solution with the lowest within-cluster sum of squares. The analysis therefore compares two independently optimized solutions rather than estimating a distribution of outcomes across individual initializations. RS42 is used as the primary solution for presenting the learned semantic hierarchy, while RS12 provides a second solution for cross-solution comparison. The resulting Level-80 clusters are treated as fine-grained candidate semantic concepts rather than predefined legal categories.

To interpret each cluster, we select the five provisions nearest to its centroid by cosine similarity. Duplicate chunks derived from the same original provision are excluded using their original identifiers. An LLM then generates a short semantic label and summary from the representative provisions. The LLM is used solely for post-hoc interpretation and does not influence the embeddings, cluster assignments, or hierarchy construction.

To obtain a multi-level semantic structure, we apply Ward hierarchical clustering with Euclidean distance to the 80 k-means centroids without additional L2 normalization.

Cutting the dendrogram at different levels yields progressively broader semantic groups.

For adjacent Ward linkage heights d_i and d_{i+1} , we calculated the successive gap $\Delta_i = d_{i+1} - d_i$. Candidate cuts were placed before prominent increases in Δ_i , producing seed-specific intermediate resolutions. A six-group resolution appeared among the retained candidate cuts for both solutions and was therefore used for side-by-side comparison.

Table 1: Composition and word-count statistics of the EU regulatory corpus.

Instrument	Provisions	Mean words	Total words
AI Act	113	373.3	42,181
Battery Regulation	96	345.4	33,157
DMA	54	378.6	20,442
DSA	93	372.1	34,603
Data Act	50	381.3	19,063
EHDS	105	324.9	34,111
ESPR	80	338.1	27,051
GDPR	99	296.8	29,384
EU Taxonomy Regulation	27	347.3	9,376
Total / overall mean	717	347.8	249,368

Note. AI Act = Artificial Intelligence Act; DMA = Digital Markets Act; DSA = Digital Services Act; EHDS = European Health Data Space Regulation; ESPR = Ecodesign for Sustainable Products Regulation; GDPR = General Data Protection Regulation. Provision and word counts are reported before chunking. The 717 retained provisions were divided into 746 units for embedding and clustering. The overall mean is calculated across the 717 retained provisions.

Table 2: Main experimental settings.

Component	Setting
Embedding	Qwen/Qwen3-Embedding-8B (4,096 dimensions); no instruction prefix; last-token pooling; L2 normalization
Chunking	717 provisions expanded to 746 embedding units
k-means	$K = 80$ (exploratory fine-grained resolution); $n_init=10$; random state 42 and 12
Hierarchy	Ward linkage with Euclidean distance, applied to the 80 k-means centroids
Hierarchy cuts	Candidate cuts selected before prominent successive increases in sorted Ward linkage heights; the shared six-group resolution used for cross-solution comparison
Interpretation	Five centroid-nearest provisions per cluster, selected by cosine similarity; duplicate chunks from the same provision excluded
LLM	gpt-5-mini; used only for post-hoc semantic labels and summaries
EuroVoc	English descriptors evaluated at the concept, Domain, and Microthesaurus levels

4.2 Cross-Solution Correspondence and Granularity

We assess global partition agreement using the Adjusted Rand Index (ARI; [Hubert and Arabie, 1985](#)) and Normalized Mutual Information (NMI; [Vinh et al., 2010](#)). Both metrics are invariant to cluster numbering.

At the concept level, we compute an 80×80 cosine-similarity matrix between the RS42 and RS12 centroids and apply the Hungarian algorithm to obtain an optimal one-to-one alignment. We additionally compare the memberships of the matched clusters using Jaccard similarity.

The Hungarian-matched cluster pairs were categorized according to their centroid cosine similar-

ity.

For descriptive analysis, pairs with similarity $s \geq 0.90$ were categorized as high-similarity pairs, those with $0.75 \leq s < 0.90$ as medium-similarity pairs, and those with $s < 0.75$ as low-similarity pairs.

The supplementary correspondence visualization displays centroid pairs with $s \geq 0.70$ to retain secondary one-to-many and many-to-one correspondences. This display threshold does not affect the matching or candidate-selection analyses.

Membership Jaccard similarity was reported separately and was not used to assign these categories.

To identify directional split and merge candidates, we examined the full centroid-similarity distribution of each cluster against the 80 clusters from the other initialization. For each cluster, we calculated the boundary margin between its highest and second-highest similarities and the normalized entropy of its softmax-normalized similarity distribution ($\tau = 0.05$).

We fixed $\tau = 0.05$ to obtain concentrated correspondence profiles and treated it as a fixed analysis parameter rather than an optimized hyperparameter.

A cluster was selected as a granularity-sensitive candidate when its boundary margin was in the lowest quartile and its normalized correspondence entropy was in the highest quartile. Candidates in the RS42-to-RS12 direction were treated as split candidates, whereas those in the reverse direction were treated as merge candidates.

These quartile-based criteria provide descriptive candidates for further inspection rather than statistically significant split or merge effects.

High-similarity pairs preserve strong centroid-

level semantic correspondence between the two solutions. Split and merge candidates have multiple similarly plausible correspondences and are therefore treated as granularity-sensitive or boundary-ambiguous. Low-similarity pairs exhibit weak centroid-level semantic correspondence.

4.3 EuroVoc Alignment Analysis

External alignment with EuroVoc. We evaluated the external semantic alignment of the learned hierarchy against EuroVoc, a manually curated multilingual thesaurus for European policy and legal information. The analysis was conducted at three EuroVoc granularities: concepts, Domains, and Microthesauri. Rather than treating EuroVoc as ground-truth cluster labels, we used it as an external reference taxonomy to examine whether the learned clusters reproduced existing policy categories or captured relations spanning multiple taxonomy categories.

Concept-level alignment. EuroVoc concept labels were encoded using the same Qwen3-Embedding-8B model used to represent the regulatory documents.

For each k-means solution, we constructed an embedding for every cluster at each hierarchical level by averaging the centroids of its member Level-80 clusters.

All learned centroids and EuroVoc concept embeddings were L2-normalized, and their cosine similarity was calculated.

For each learned cluster, we retained the 20 most similar EuroVoc concepts. Concept-level alignment was summarized using the highest similarity, the mean similarity of the top five concepts, and the minimum similarity among the top five concepts. The top-20 similarities were converted into normalized concept weights using a temperature-scaled softmax. These weights were subsequently aggregated to EuroVoc Domains and Microthesauri.

We compute cosine similarity using Equation 1 and normalize the top-20 similarities using Equation 2.

The cosine similarity between a learned cluster centroid c and a EuroVoc concept embedding e_j is defined as

$$s(c, e_j) = \frac{c^\top e_j}{\|c\|_2 \|e_j\|_2}. \quad (1)$$

The softmax-normalized weight assigned to concept j is

$$w_j = \frac{\exp(s(c, e_j)/\tau)}{\sum_{k \in \text{Top}_{20}(c)} \exp(s(c, e_k)/\tau)}. \quad (2)$$

where τ is the temperature parameter.

For consistency with the cross-solution analysis, we used the same fixed temperature, $\tau = 0.05$, and retained the top 20 concepts for each learned cluster.

Domain-level alignment. Each selected EuroVoc concept was mapped to its associated EuroVoc Domain. When a concept belonged to multiple Domains, its concept weight was divided equally among them to prevent duplicated weight allocation. Domain weights were then summed and normalized within each learned cluster to form a Domain alignment profile.

Candidate-pattern analysis. The three continuous measures—concept similarity, Domain purity, and normalized Domain entropy—were used as the primary alignment results. For qualitative candidate selection, we additionally defined four descriptive patterns using within-granularity tertiles. Type A represented high similarity, high purity, and low entropy; Type B represented high similarity, low purity, and high entropy; Type C represented low similarity, low purity, and high entropy; and Type D contained the remaining mixed cases. These categories were used only to identify clusters for further inspection and were not treated as ground-truth classes.

Microthesaurus robustness analysis. To test whether apparent Domain crossing was an artifact of coarse Domain categories, we reaggregated the same concept weights to EuroVoc Microthesauri. We did not independently embed Microthesaurus labels; therefore, the comparison preserved the original concept-level evidence. Microthesaurus purity and normalized entropy were calculated analogously to the Domain metrics.

Type B clusters that remained relatively low in Microthesaurus purity and high in Microthesaurus entropy were retained as persistent taxonomy-crossing candidates. Type C clusters showing the same pattern were categorized as weak-EuroVoc-explanation candidates.

Candidate validation. Persistent candidates were further evaluated through hierarchical deduplication, cross-seed distributional matching, and sensitivity analyses. Nested candidates sharing at least

80% containment of their Level-80 members were treated as the same semantic branch. Cross-seed recurrence was measured using weighted Jaccard similarity between concept-weight distributions and Jensen–Shannon similarity between Domain and Microthesaurus profiles. The cross-seed correspondence score was calculated as the arithmetic mean of these three similarity measures. Robustness was evaluated across top-k settings of 10, 15, and 20 concepts and across three quantile-threshold settings. Jensen–Shannon similarity is defined as one minus Jensen–Shannon divergence computed using base-2 logarithms (see Appendix E).

Experimental setting. Table 2 summarizes the main experimental configuration.

5 Results

5.1 RQ1: Discovered Semantic Structures

The clustering produced 80 fine-grained candidate concepts. Representative provisions associated these clusters with regulatory subjects, actors, obligations, and procedures, including high-risk AI conformity obligations, battery product requirements, oversight of notified bodies, data-access mechanisms, regulatory reporting, and investigation and enforcement procedures.

Ward clustering organized the 80 base clusters into six broader semantic groups for each solution as shown in Figure 1 and Table 3. Both solutions exhibited recurring themes related to AI, batteries, data governance, and regulatory oversight, alongside cross-regulatory functional groups. Their hierarchical organization differed: RS42 distinguished digital and data regulatory governance from data access and sharing, whereas RS12 contained two broader cross-domain groups emphasizing regulatory implementation and enforcement.

For RS42, the six groups contained 28, 15, 6, 16, 6, and 9 base clusters, respectively. They were labelled as legislative framing and cross-references; battery governance and compliance; oversight of notified bodies; AI-system compliance and supervision; digital and data regulatory enforcement; and data access and sharing.

RS12 produced six groups containing 21, 23, 6, 18, 7, and 5 base clusters, respectively, with different internal branch configurations.

The induced hierarchies exhibited both domain-oriented groups combining regulatory subjects and functions and cross-domain groups organized around regulatory functions.

5.2 RQ2: Cross-Solution Correspondence and Granularity

The RS42 and RS12 partitions achieved an ARI of 0.520 and an NMI of 0.830. Hungarian matching between the 80 clusters produced a mean matched-centroid cosine similarity of 0.943 and a median similarity of 0.966. The mean membership Jaccard similarity was 0.492.

Based on the centroid-similarity thresholds, 64 of the 80 Hungarian-matched cluster pairs (80.0%) were categorized as high-similarity, 14 (17.5%) as medium-similarity, and two (2.5%) as low-similarity pairs.

The correspondence-distribution analysis identified 10 split and nine merge candidates with relatively small boundary margins and high normalized correspondence entropy. These candidates exhibited multiple similarly plausible centroid-level correspondences, suggesting potential differences in local semantic boundaries or granularity rather than confirmed membership-based splitting or merging.

The complete cross-seed correspondence among all 80 base clusters is provided in Appendix G.

The observed difference between partition agreement and centroid-level correspondence indicates that similar cluster representations may be associated with different sets of embedding units across the two solutions.

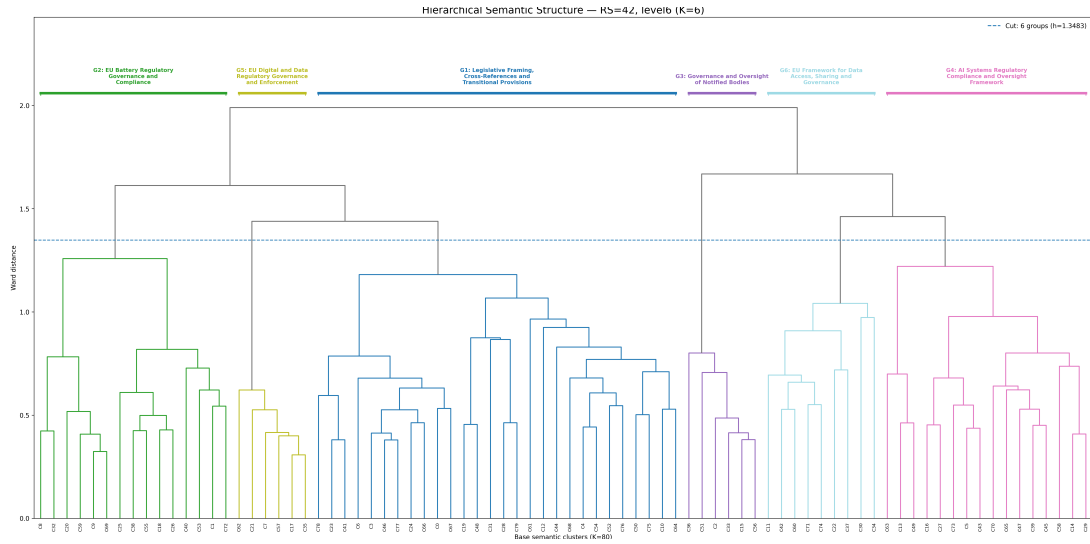
5.3 RQ3: EuroVoc Alignment

Mean top-five EuroVoc concept similarity increased from 0.643 at the 28-cluster level to 0.663 at the six-cluster level for RS12, and from 0.643 at the 23-cluster level to 0.678 at the six-cluster level for RS42. At the shared six-cluster level, Domain purity was 0.680 for RS12 and 0.669 for RS42, while normalized Domain entropy was 0.301 and 0.303, respectively.

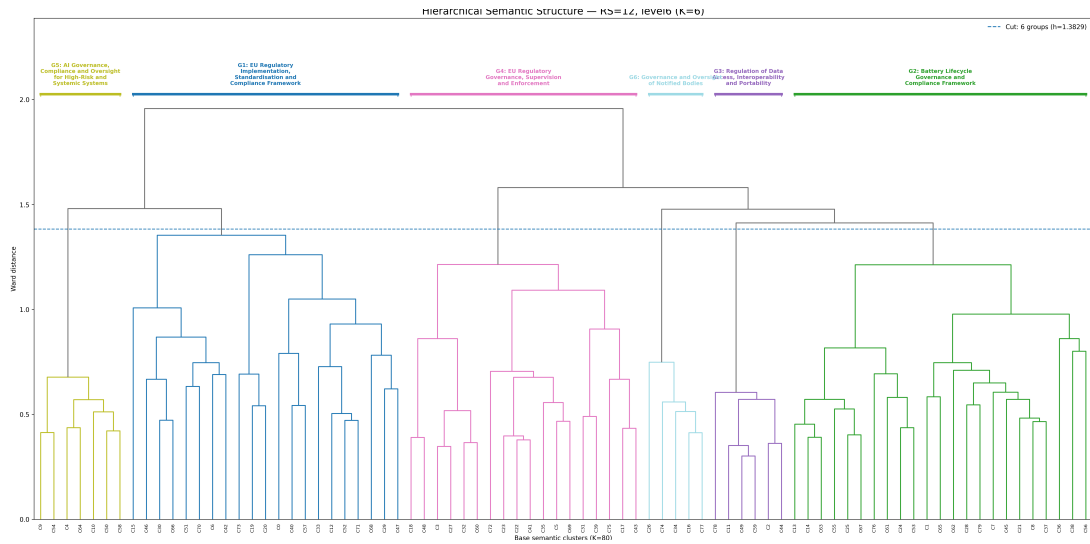
Excluding the single root cluster from each initialization, the analysis covered 114 cluster-level instances.

Of these, 27 were classified as Type A, four as Type B, 20 as Type C, and 63 as Type D.

Three of the four Type B instances also exhibited relatively low purity and high entropy at the Microthesaurus level. After hierarchical deduplication, these instances corresponded to two independent branches. Their cross-seed correspondence scores ranged from 0.831 to 0.864, and their robustness across top-k and threshold settings was moderate.



(a) RS42



(b) RS12

Figure 1: Six-group hierarchical semantic structures obtained from RS42 and RS12. Colored branches represent Level-6 semantic groups, while gray branches connect different groups. Colors are aligned across panels according to semantic correspondence rather than group identifiers.

Table 3: Level-6 semantic groups for RS42 and RS12. Group identifiers are initialization-specific; the number of constituent Level-80 clusters is shown.

RS	Group	Domain	Semantic label	# Base clusters
RS42: Level-6 hierarchy				
42	G1	Cross-domain	Legislative Framing, Cross-References and Transitional Provisions	28
42	G2	Battery	EU Battery Regulatory Governance and Compliance	15
42	G3	Product compliance	Governance and Oversight of Notified Bodies	6
42	G4	AI	AI Systems Regulatory Compliance and Oversight Framework	16
42	G5	Digital and data	EU Digital and Data Regulatory Governance and Enforcement	6
42	G6	Data	EU Framework for Data Access, Sharing and Governance	9
RS12: Level-6 hierarchy				
12	G1	Cross-domain	EU Regulatory Implementation, Standardisation and Compliance Framework	21
12	G2	Battery	Battery Lifecycle Governance and Compliance Framework	23
12	G3	Data	Regulation of Data Access, Interoperability and Portability	6
12	G4	Cross-domain	EU Regulatory Governance, Supervision and Enforcement	18
12	G5	AI	AI Governance, Compliance and Oversight for High-Risk and Systemic Systems	7
12	G6	Product compliance	Governance and Oversight of Notified Bodies	5

6 Discussion

6.1 Mixed Domain and Regulatory-Function Organization

The discovered hierarchy did not follow a single organizing principle. Some branches were primarily associated with regulatory domains, such as AI systems and batteries, whereas others grouped provisions according to cross-cutting functions, including supervision, enforcement, compliance, and data access. The embedding space therefore appears to organize provisions through a mixture of domain and regulatory-function information.

This mixed organization is important because individual provisions often express both what is regulated and how regulation is implemented. The learned hierarchy may therefore complement domain-based legal taxonomies by making regulatory functions across different instruments more visible.

The differences between RS42 and RS12 further suggest that domain-specific and cross-regulatory functional themes may be organized differently across alternative clustering solutions.

6.2 Broad Cross-Solution Correspondence and Local Granularity Sensitivity

While RQ1 identified recurring regulatory themes across the two induced hierarchies, their differences in cluster composition and branch organization motivate a closer examination of cross-solution correspondence.

The cross-solution analysis reveals a distinction between partition agreement and centroid-level semantic correspondence. Although the two solutions showed substantial centroid similarity, their membership overlap was lower, indicating that semantically similar cluster centroids may be associated with different sets of provisions.

The granularity-sensitive candidates provide a more localized view of these differences. Their correspondence profiles show that some clusters have multiple similarly plausible centroid-level matches across solutions. Such patterns may reflect differences in semantic boundaries or granularity, although centroid similarity alone cannot establish whether a concept has been split or merged.

These findings suggest that broadly corresponding semantic structures can coexist with local differences in cluster membership and potentially in semantic granularity. However, centroid-level correspondence does not establish actual splitting or

merging, nor does it directly measure intra-cluster semantic coherence. Further examination of cluster memberships and representative provisions is required to determine the semantic nature of the identified candidates.

6.3 Partial EuroVoc Alignment Rather Than Taxonomy Reproduction

Mean EuroVoc concept similarity was higher at the six-cluster level than at the reported finer-grained levels for both solutions. This does not necessarily imply that broader clusters align with a single EuroVoc Domain.

This difference suggests that EuroVoc and the learned hierarchy emphasize different aspects of legal semantics. EuroVoc primarily provides a curated policy-subject organization, whereas the learned structure groups provisions according to semantic proximity in the regulatory corpus. Partial alignment is therefore expected, but complete taxonomy reproduction is neither observed nor required.

6.4 Cross-Taxonomy Relations and Scope of Evidence

The two persistent candidate branches combined concepts related to EU law, environmental policy, and technology-oriented regulation, including extended producer responsibility, ecodesign, conformity marking, and delegated regulation. Their potential contribution lies not in identifying concepts absent from EuroVoc, but in grouping existing concepts along a regulatory-functional relation that spans curated taxonomy boundaries.

Only two independent branches remained after hierarchical deduplication, and their robustness across analysis settings was moderate. We therefore regard them as qualitative complementary-structure candidates rather than confirmed novel semantic categories. Stronger validation requires expert assessment, article-level analysis, and external baselines matched for semantic similarity and hierarchical level.

7 Conclusion

We identified recurring regulatory semantic themes and differences in hierarchical organization across two K-means solutions. Cross-solution analysis revealed broad centroid-level semantic correspondence despite differences in cluster membership, alongside candidate split–merge patterns suggesting potential local granularity differences.

Post-hoc EuroVoc alignment showed partial agreement with established categories and identified two qualitative cross-taxonomy candidate branches.

Future work will examine additional random initializations and finer hierarchical resolutions, using PCA, probabilistic mixture models, and information-geometric approaches to investigate the relationships among assignment variability, semantic coherence, and granularity. Beyond coherent cluster interiors, we will examine semantic boundary regions and granularity-sensitive branches where cross-solution correspondence becomes ambiguous and relationships among multiple regulatory concepts may emerge. This may provide a more fine-grained understanding of semantic organization in regulatory texts.

Limitations

The corpus is limited to 717 English-language provisions from nine EU regulations, restricting generalizability to other legal systems, document types, and languages.

The selected regulations do not constitute a representative sample of the EU legal framework as a whole.

The cross-solution analysis compares only two independently optimized K-means solutions. Because each solution was selected from 10 internal initializations, the comparison does not provide a distributional estimate of stability across random initializations.

The choice of $K = 80$, the hierarchy-cut criteria, and the fixed temperature $\tau = 0.05$ may affect the resulting semantic structures and correspondence profiles. These settings were treated as exploratory design choices rather than optimized hyperparameters.

Centroid-level correspondence does not establish identical cluster membership, semantic inclusion, or actual splitting and merging. Neither intra-cluster semantic coherence nor downstream retrieval effectiveness was evaluated. Further validation requires additional random initializations, cross-solution membership analysis, and independent semantic assessment.

Ethics Statement

This study analyzes publicly available EU regulatory texts obtained from EUR-Lex. The dataset consists of 717 provisions from nine EU regula-

tions and does not include private legal documents or personal user data.

The proposed framework is intended for exploratory semantic analysis rather than automated legal interpretation or decision-making. LLM-generated cluster labels and summaries are used only for post-hoc interpretation and do not constitute independent validation of legal meaning.

The discovered semantic structures and their alignment with EuroVoc may reflect limitations or biases in the embedding model, the regulatory corpus, and the interpretation process. In particular, semantic similarity between regulatory provisions does not necessarily imply equivalent legal meaning, obligations, or applicability.

Consequently, the resulting clusters, labels, and cross-regulatory relations should not be treated as authoritative legal classifications or used as the sole basis for legal or compliance decisions. Any practical application would require appropriate human review and validation by qualified domain experts.

Acknowledgments

I would like to express my sincere gratitude to Professor Tomonari Masada of Rikkyo University for his guidance and valuable insights during my master’s research, which laid the foundation for this study.

I would like to thank Honda Motor Co., Ltd. for providing computational resources and supporting this research.

I also thank the anonymous reviewers of NLLP 2026 for their valuable feedback and suggestions to improve this manuscript.

References

- Andrei-Marius Avram, Vasile Pais, and Dan Ioan Tufis. 2021. [PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101. INCOMA Ltd.
- Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. [Large-scale multi-label text classification on EU legislation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. [MultiEURLEX: A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer](#). In *Proceedings of*

the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6974–6996. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Maarten Grootendorst. 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). *arXiv preprint arXiv:2203.05794*.

Christian Hennig. 2007. [Cluster-wise assessment of cluster stability](#). *Computational Statistics & Data Analysis*, 52(1):258–271.

Lawrence Hubert and Phipps Arabie. 1985. [Comparing partitions](#). *Journal of Classification*, 2(1):193–218.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. [Self-supervised topic taxonomy discovery in the box embedding space](#). *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Amin Hosseiny Marani and Eric P. S. Baumer. 2023. [A review of stability in topic modeling: Metrics for assessing and techniques for improving stability](#). *ACM Computing Surveys*, 56(5):1–32.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. [Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance](#). *Journal of Machine Learning Research*, 11:2837–2854.

Ulrike von Luxburg. 2010. [Clustering stability: An overview](#). *Foundations and Trends in Machine Learning*, 2(3):235–274.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.

A Implementation and Reproducibility Details

Table 4 reports additional implementation details for the clustering, hierarchical analysis, cross-solution matching, LLM-based interpretation, and EuroVoc alignment experiments.

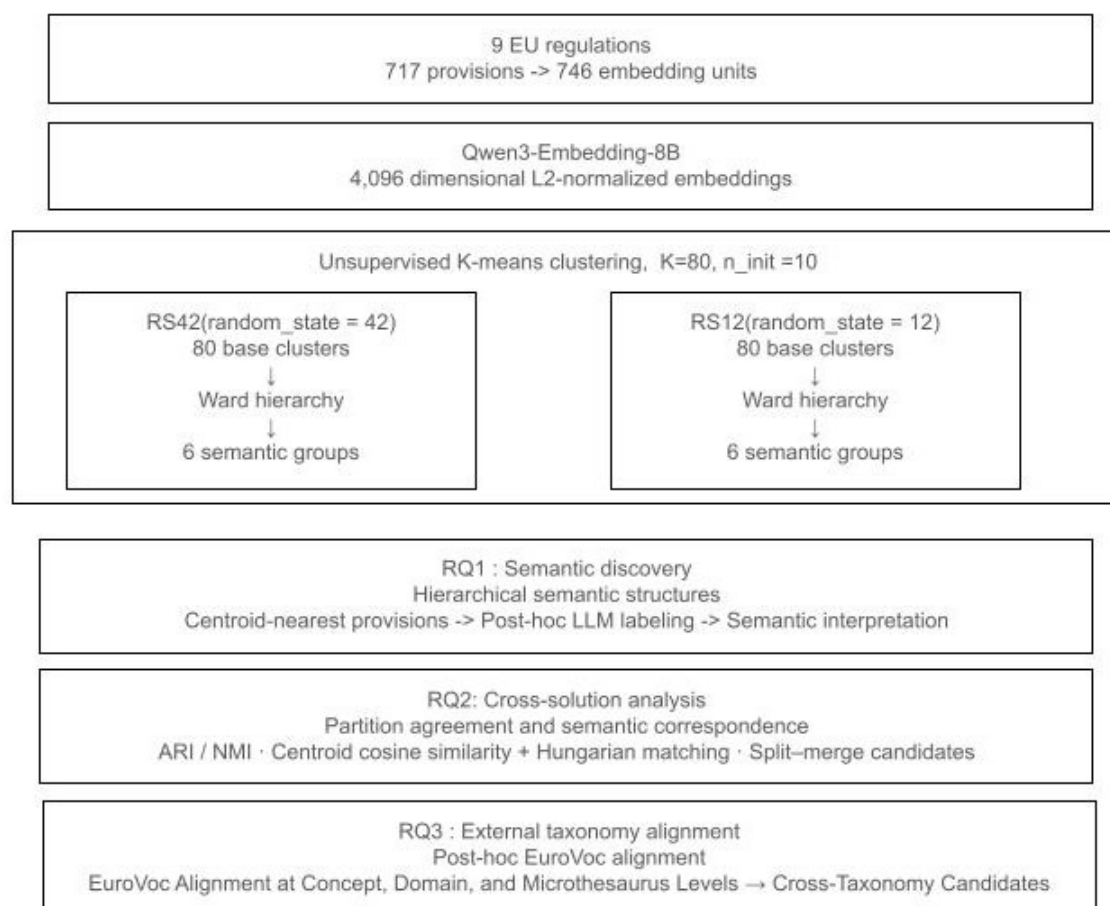


Figure 2: Overview of the proposed framework. The same regulatory embeddings are used to construct two independently optimized K-means solutions and their Ward hierarchies. The resulting structures are examined through hierarchical semantic discovery (RQ1), cross-solution correspondence analysis (RQ2), and post-hoc EuroVoc alignment (RQ3). LLM-generated labels and EuroVoc are not used during clustering.

Table 4: Implementation and reproducibility details.

Component	Implementation	Configuration
Software libraries	Python; NumPy; pandas; Sentence-Transformers; scikit-learn; SciPy	Python 3.12.13; NumPy 2.0.2; pandas 2.2.2; Sentence-Transformers 5.6.0; scikit-learn 1.6.1; SciPy 1.16.3.
k-means clustering	<code>sklearn.cluster.KMeans</code>	$K = 80$; $n_init=10$; $random_state \in \{42, 12\}$. For each random state, 10 internal initializations were performed, and the solution with the lowest within-cluster sum of squares was retained. Initialization: k-means++; maximum iterations: 300; tolerance: 10^{-4} ; algorithm: lloyd.
Ward hierarchical clustering	<code>scipy.cluster.hierarchy.linkage</code>	Ward linkage with Euclidean distance, applied to the 80 K-means centroids without additional L2 normalization. Candidate hierarchy cuts were selected before prominent successive increases in the sorted Ward linkage heights.
Hungarian matching	<code>scipy.optimize.linear_sum_assignment</code>	An 80×80 cosine-similarity matrix was constructed between the RS42 and RS12 centroids. The Hungarian algorithm was applied to the negative similarity matrix to obtain the one-to-one assignment maximizing total centroid cosine similarity.
LLM interpretation	OpenAI gpt-5-mini	Five centroid-nearest provisions were used for post-hoc cluster labeling and summarization. Each representative provision was truncated to a maximum of 2,000 characters. The decoding temperature was set to 0.0, and prompt version <code>cluster_label_v2.0</code> was used. Other decoding parameters were not explicitly specified. The generated outputs did not affect the embeddings, cluster assignments, or hierarchy construction.
EuroVoc	EU Publications Office	English descriptors and hierarchy data retrieved through the EU Publications Office SPARQL endpoint on 10 August 2026 (EuroVoc graph 100141). No fixed release identifier was recorded.

B Prompt for Cluster Interpretation

The following prompt was used to generate post-hoc semantic labels and summaries from the five representative provisions nearest to each cluster centroid. The generated outputs were used only for interpretation and did not affect the embeddings, cluster assignments, or hierarchy.

You are an expert in legal NLP, semantic clustering, and European digital regulations.

Your task is to identify the central legal concept represented by a semantic cluster of legal provisions.

The input consists of representative legal articles selected from the centroid of one semantic cluster.

Please identify the shared legal meaning across the articles rather than summarizing each article independently.

Requirements:

1. The label must represent the common legal concept.
2. The label should be a concise noun phrase.
3. Do not use sentence-style labels.
4. Prefer abstract legal concepts rather than document titles.
5. The summary should explain:
 - the central legal concept,
 - the regulated actors,
 - the main rights, obligations, and procedures,
 - how this cluster differs from related legal concepts.
6. Do not simply list keywords.
7. Use only the information contained in the representative articles.

8. Return the result as valid JSON only.

Label Requirements:

- The English label must represent the shared legal concept.
- The English label must be a noun phrase containing 2 to 8 words.
- Avoid full sentences.
- Avoid overly generic labels such as Miscellaneous, Data, AI, or Rights.
- Prefer concept-level labels such as User Data Access Rights, AI Risk Management, Environmental Sustainability Requirements, or Digital Market Competition Rules.

Summary Requirements:

- The English and Japanese summaries should each cover:
1. The central legal concept.
 2. The regulated actors.
 3. The main rights, obligations, and procedures.
 4. The distinction from related semantic clusters.

Key Concept Requirements:

- Return between 3 and 6 key concepts.
- Each concept must contain 1 to 4 words.
- Use legal or regulatory concepts.
- Avoid document titles.
- Avoid repeating the cluster label.

Confidence Requirement:

Provide a confidence score between 0 and 1. The score represents how consistently the representative articles describe one shared legal concept. It does not represent factual correctness.

Semantic Scope:

Describe the concept range covered by the cluster.

Semantic Boundary:

State briefly what the cluster includes and excludes.

Do not claim exclusions that cannot reasonably be inferred from the supplied representative articles.

C Chunking Configuration and Input-Length Statistics

The Qwen3-Embedding-8B tokenizer was used to measure input lengths. Provisions exceeding 1,500 tokens were segmented by prioritizing legal-structure boundaries, with a target chunk length of 1,024 tokens and a maximum of 1,500 tokens.

A token-based fallback with 128-token overlap was available for units that could not be segmented within the maximum length using legal-structure boundaries.

Of the 717 original provisions, 16 were split, producing 746 embedding units. The maximum original provision length was 6,199 tokens. The resulting embedding units had a mean length of 423.47 tokens, a median of 333.50 tokens, and an interquartile range of 158.25–607.75 tokens. Their lengths ranged from 8 to 1,449 tokens.

D Evaluation Metrics and Candidate Selection

This appendix provides the mathematical definitions of the evaluation metrics and candidate-selection criteria used in the cross-solution analysis described in Section 4.2.

Let $X = \{x_1, \dots, x_N\}$ denote the set of $N = 746$ embedding units, and let U and V denote the cluster assignments obtained from RS42 and RS12, respectively. Both solutions contain $K = 80$ clusters.

D.1 Global Partition Agreement

We evaluate global partition agreement using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

Adjusted Rand Index. ARI measures agreement between two partitions by comparing pairwise co-membership, adjusted for chance:

$$\text{ARI}(U, V) = \frac{\text{RI}(U, V) - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}, \quad (3)$$

where RI denotes the Rand Index.

Normalized Mutual Information. NMI measures the shared information between two cluster assignments:

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)}, \quad (4)$$

where $I(U; V)$ denotes mutual information, and $H(U)$ and $H(V)$ denote the entropies of the respective cluster assignments.

Both metrics are invariant to permutations of cluster identifiers.

The analysis uses the same 746 embedding units in both solutions, aligned by their unique identifiers.

D.2 Centroid Correspondence and Membership Overlap

Centroid cosine similarity. Let $\mathbf{c}_i^{42}$ and $\mathbf{c}_j^{12}$ denote the centroids of clusters i and j in the RS42 and RS12 solutions, respectively.

Their cosine similarity is defined as

$$S_{ij} = \frac{(\mathbf{c}_i^{42})^\top \mathbf{c}_j^{12}}{\|\mathbf{c}_i^{42}\|_2 \|\mathbf{c}_j^{12}\|_2}. \quad (5)$$

We construct an 80×80 cosine-similarity matrix to evaluate cross-solution centroid correspondence.

Hungarian matching. An optimal one-to-one correspondence between the two sets of centroids is obtained by solving

$$\pi^* = \arg \max_{\pi \in \mathcal{S}_K} \sum_{i=1}^K S_{i, \pi(i)}, \quad (6)$$

where $\mathcal{S}_K$ denotes the set of permutations of $K = 80$ cluster indices.

The mean matched-centroid similarity is

$$\bar{S}_{\text{matched}} = \frac{1}{K} \sum_{i=1}^K S_{i, \pi^*(i)}. \quad (7)$$

Membership Jaccard similarity. Let A_i and $B_{\pi^*(i)}$ denote the sets of embedding-unit identifiers assigned to a matched pair of clusters.

Their membership overlap is defined as

$$J_i = \frac{|A_i \cap B_{\pi^*(i)}|}{|A_i \cup B_{\pi^*(i)}|}. \quad (8)$$

The mean membership overlap is

$$\bar{J} = \frac{1}{K} \sum_{i=1}^K J_i. \quad (9)$$

Centroid similarity measures the proximity of cluster representatives in the embedding space, whereas Jaccard similarity measures agreement in cluster membership.

Hungarian matching is performed using centroid cosine similarity. Membership Jaccard similarity is calculated separately and does not affect the matching.

D.3 Centroid Similarity Categories

For descriptive analysis, Hungarian-matched cluster pairs are categorized according to their centroid cosine similarity s :

$$\text{Category}(s) = \begin{cases} \text{High}, & s \geq 0.90, \\ \text{Medium}, & 0.75 \leq s < 0.90, \\ \text{Low}, & s < 0.75. \end{cases} \quad (10)$$

These thresholds describe centroid-level correspondence and do not establish general clustering stability or semantic coherence.

D.4 Granularity-Sensitive Correspondence Metrics

For each cluster, we examine its cosine similarities to all $K = 80$ clusters in the other solution.

Let

$$s_{(1)} \geq s_{(2)} \geq \dots \geq s_{(K)} \quad (11)$$

denote the sorted similarities.

Boundary margin. The boundary margin is defined as

$$M = s_{(1)} - s_{(2)}. \quad (12)$$

A small margin indicates that the two highest-ranked centroid correspondences have similar cosine-similarity values.

Softmax-normalized correspondence weights. We transform the centroid similarities into a normalized distribution:

$$p_j = \frac{\exp(s_j/\tau)}{\sum_{k=1}^K \exp(s_k/\tau)}, \quad (13)$$

where the temperature parameter is fixed at $\tau = 0.05$.

Normalized correspondence entropy. The entropy of the correspondence distribution is defined as

$$H(p) = - \sum_{j=1}^K p_j \log p_j. \quad (14)$$

We normalize the entropy by its maximum value for K possible correspondence targets:

$$H_{\text{norm}}(p) = \frac{H(p)}{\log K}. \quad (15)$$

Higher normalized correspondence entropy indicates a more dispersed similarity-based correspondence profile.

This measure describes the distribution of centroid similarities and should not be interpreted as a direct measure of uncertainty about legal meaning.

Effective number of matches. As an auxiliary descriptive measure, we compute

$$N_{\text{eff}} = \exp(H(p)). \quad (16)$$

This quantity represents the entropy-equivalent number of correspondence targets under the softmax distribution.

Granularity score. An additional auxiliary score is calculated from the relative similarities of the second- through fifth-ranked correspondence targets:

$$G = \frac{1}{4} \sum_{r=2}^5 \frac{s_{(r)}}{s_{(1)} + \epsilon}, \quad (17)$$

where $\epsilon = 10^{-12}$.

The score summarizes how closely the next four correspondence similarities approach the highest similarity.

Effective matches and the granularity score are auxiliary descriptive measures and are not used in the final split/merge candidate-selection rule.

D.5 Directional Split and Merge Candidate Selection

We identify granularity-sensitive candidates using the boundary margin and normalized correspondence entropy.

For each direction, a cluster is selected when both conditions hold:

$$M_i \leq Q_{0.25}(M) \quad (18)$$

and

$$H_{\text{norm},i} \geq Q_{0.75}(H_{\text{norm}}), \quad (19)$$

where $Q_{0.25}$ and $Q_{0.75}$ denote the first and third quartiles of the corresponding metric distributions.

The criteria are applied independently in both directions.

Candidates identified in the RS42-to-RS12 direction are referred to as split candidates, whereas candidates identified in the RS12-to-RS42 direction are referred to as merge candidates.

These criteria identify clusters with multiple similarly plausible centroid correspondences.

The resulting candidates are descriptive indicators of potential granularity sensitivity or correspondence ambiguity.

They do not establish actual membership-based splitting or merging, statistically significant granularity differences, or general clustering stability across random initializations.

E Jensen–Shannon Similarity

Let P and Q denote two normalized EuroVoc Domain or Microthesaurus alignment profiles, defined over the union of their category identifiers.

We define their mixture distribution as

$$M = \frac{P + Q}{2}. \quad (20)$$

The Jensen–Shannon divergence is

$$\text{JSD}(P, Q) = \frac{1}{2} D_{\text{KL}}(P \| M) + \frac{1}{2} D_{\text{KL}}(Q \| M), \quad (21)$$

where

$$D_{\text{KL}}(P \| M) = \sum_{i: P_i > 0} P_i \log_2 \frac{P_i}{M_i}. \quad (22)$$

Following the implementation, we convert divergence into similarity as

$$S_{\text{JS}}(P, Q) = 1 - \text{JSD}(P, Q). \quad (23)$$

The similarity ranges from 0 to 1, with higher values indicating greater agreement between the two alignment profiles.

F EuroVoc Alignment Metrics

For each learned cluster c , EuroVoc concept weights were aggregated at the Domain level and normalized to form a Domain alignment profile. Let $S_{c,d}$ denote the aggregated weight assigned to

Type	Similarity	Purity	Entropy
A	High	High	Low
B	High	Low	High
C	Low	Low	High
D	Other	Other	Other

Table 5: Descriptive EuroVoc alignment patterns. High and low values are defined using within-granularity upper and lower tertiles, respectively. Type D contains cases that do not satisfy the criteria for Types A–C.

Domain d . The normalized Domain weight is defined as

$$P_D(d | c) = \frac{S_{c,d}}{\sum_{d'} S_{c,d'}}. \quad (24)$$

Domain purity measures the concentration of the alignment profile on its most strongly represented Domain:

$$\text{Purity}_D(c) = \max_d P_D(d | c). \quad (25)$$

Normalized Domain entropy measures the dispersion of the alignment profile across Domains:

$$H_{D,\text{norm}}(c) = \frac{-\sum_d P_D(d | c) \log P_D(d | c)}{\log D}, \quad (26)$$

where D denotes the number of EuroVoc Domains represented in the corresponding alignment profile. Lower entropy indicates concentration on a smaller number of Domains, whereas higher entropy indicates a more distributed Domain profile.

For qualitative analysis, clusters were categorized into four descriptive alignment patterns using within-granularity tertiles of mean top-five EuroVoc concept similarity, Domain purity, and normalized Domain entropy. Table 5 summarizes the criteria.

G Additional Hierarchy Visualizations and Cross-Solution Correspondence

Figures 3 and 4 show additional intermediate hierarchy cuts for RS42 and RS12, respectively. These visualizations complement the shared six-group resolution reported in Section 5.1.

Figure 5 visualizes cross-solution centroid correspondences among all 80 base clusters. It supplements the aggregate partition agreement, matched-centroid similarity, and membership-overlap results reported in Section 5.2.

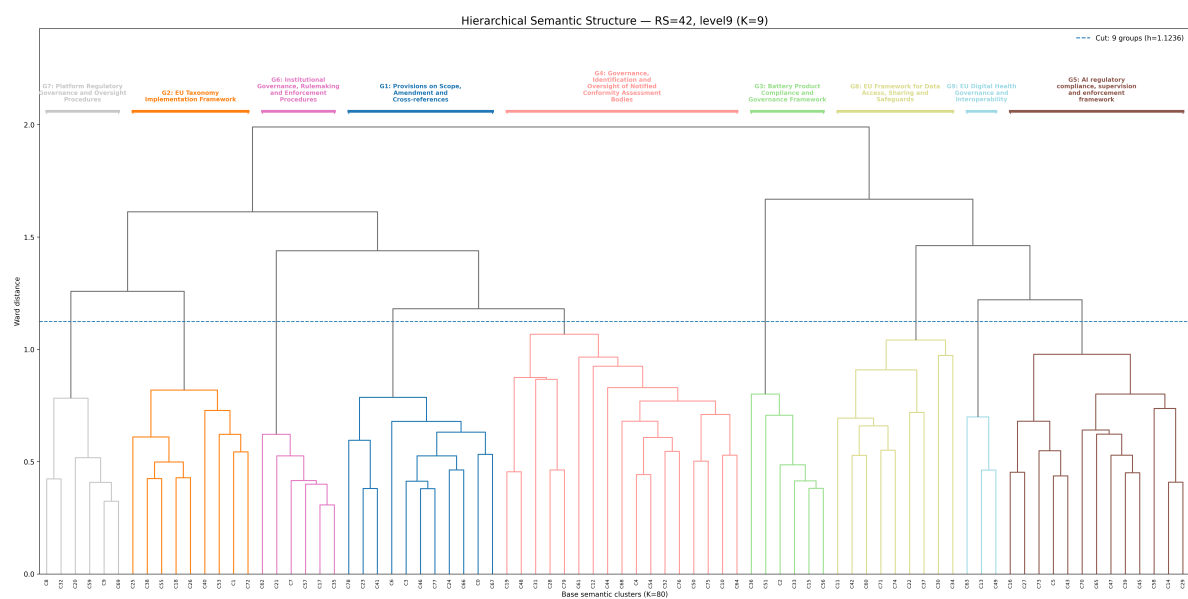


Figure 3: Dendrogram showing the nine-group intermediate hierarchical structure for RS42. Colored branches represent the nine groups at the selected cut.

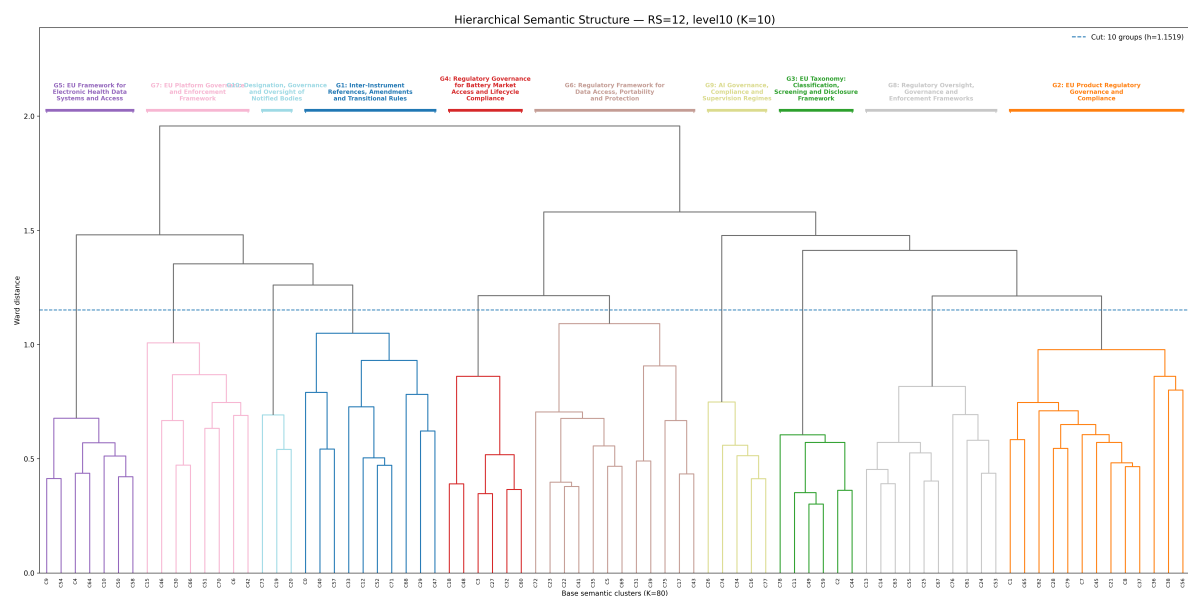


Figure 4: Dendrogram showing the ten-group intermediate hierarchical structure for RS12. Colored branches represent the ten groups at the selected cut.

Cross-seed Semantic Correspondence RS42 → RS12

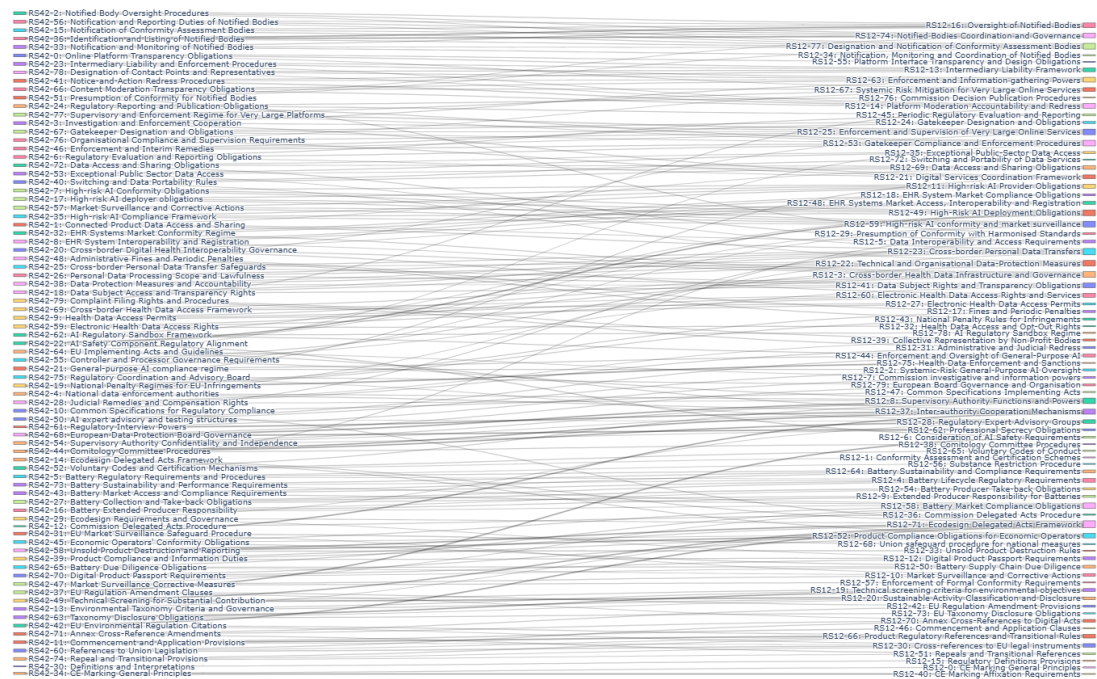


Figure 5: Cross-solution semantic correspondence between the 80 base clusters obtained with RS42 and RS12. Lines connect centroid pairs with cosine similarity $s \geq 0.70$. This display threshold retains secondary correspondences potentially associated with one-to-many and many-to-one patterns and is not used in the matching or candidate-selection analyses.