Tracing Definitions: Lessons from Alliance Contracts in the Biopharmaceutical Industry

Maximilian Kreutner[1], Doerte Leusmann[2], Florian Lemmerich[2], Carolin Haeussler[2]
[1]University of Mannheim, [2]University of Passau
Correspondence: maximilian.kreutner@uni-mannheim.de; doerte.leusmann@uni-passau.de;
florian.lemmerich@uni-passau.de; carolin.haeussler@uni-passau.de

Abstract

Definitions in alliance contracts play a critical role in shaping agreements, yet they can also lead to costly misunderstandings. This is exemplified by the multimillion-dollar AstraZeneca-Euopean Commission (EC) dispute, where the interpretation of 'best reasonable effort' became the focal point of contention. In this interdisciplinary study, we leverage natural language processing (NLP) to systematically analyze patterns in the definitions included in alliance contracts. More specifically, we categorize the content of definitions into topics, identify common terms versus outliers that are semantically dissimilar and infrequently used, and track how definitions evolve over time. Analyzing a dataset of 380,131 definitions from 12,468 alliance contracts in the biopharmaceutical industry, we distinguish that definitions span legal, technological, and social topics, with social terms showing the highest dissimilarity across contracts. Using dynamic topic modeling, we explore how the content of definitions has shifted over two decades (2000–2020) and identify prevalent trends suggesting that contractual definitions reflect broader economic contexts. Notably, our results reveal that the AstraZeneca-EC dispute arose from an outlier—a highly unusual definition—that could have been flagged using NLP. Overall, these findings highlight the potential of data-driven approaches to uncover patterns in alliance contracts.

1 Introduction

Collaboration between firms is a crucial building block for a globalized economy. Contracts, i.e., legal agreements that determine promises, obligations, and the future course of action (Macneil, 1978), constitute the backbone of alliances as organizational forms of collaboration. A key element within such legal documents that shapes the interpretation of the contract is definitions, i.e., exact descriptions of key terms used in the contract. The crucial function of definitions—and their potential to create costly misunderstandings—became evident to a broader audience in August 2020, when the European Commission and AstraZeneca included a contractual definition of AstraZeneca's 'best reasonable efforts' for vaccine supply. Despite this, the alliance partners apparently did not agree on what the term meant, as this became the subject of a long and costly high-profile legal dispute halting the collaboration for nearly half a year. After the litigation had escalated, the partners eventually "[...] have been able to reach a common understanding which allows [us] to move forward and work in collaboration [...]". [1]

Given that these contracts are texts that constitute the foundation of alliances, they present a complex and unique research setting at the intersection of NLP, data science, and economics. Thus, in this paper, we set out to study a large corpus of alliance contracts through a unique lens, i.e., by tracing and analyzing the definitions used. In doing so, we can automatically find unusual — and therefore potentially risky — definitions of terms, identify topical areas covered by definitions in alliance contracts, and can observe trends in terms defined in the contracts over time.

In this paper, we study a corpus of 12,468 alliance contracts from the biopharmaceutical industry. We begin by extracting texts—specifically definitions—from these contracts, comparing different automated methods in the process. This extraction process resulted in a corpus of 380,131 definitions. By analyzing the similarity of definitions using embeddings, we then demonstrate, for instance, that the disputed 'best reasonable effort' definition in the AstraZeneca-EC case was highly dissimilar and unusual. Using topic modeling, we identify

[1]Ruud Dobber, AstraZeneca senior executive, https://www.reuters.com/world/europe/astrazeneca-eu-reach-settlement-delivery-covid-19-vaccine-doses-2021-09-03/, accessed on 15.01.2025

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 1–15

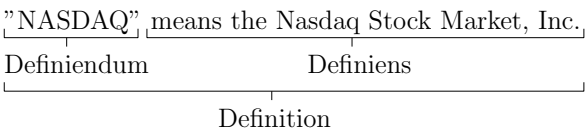

Figure 1: Definiendum and Definiens of a definition.

key content categories of definitions. To further trace trends in definitions over time, we leverage our dataset's longitudinal design, which includes alliance contracts signed between 2000 and 2020.

Our main contribution is the application of NLP methods to real-world contract texts. In so doing, we demonstrate accurate and efficient methods for extracting and analyzing contract texts. Furthermore, we contribute by providing a comprehensive dataset—a definition corpus. Beyond these methodological and data-related contributions, we advance the literature in two ways: First, our findings reveal that legal definitions are the most similar across alliance contracts, whereas technological and social definitions show more variation, making them more dissimilar, unusual, and specific to each alliance. Therefore, alliance partners should pay special attention to those definitions. Second, we demonstrate that external events, such as the COVID-19 pandemic and financial market shocks, coincide with shifts in the usage and the content of contract definitions. This finding expands the traditional view of contracts as predominantly static legal instruments, instead emphasizing their role as adaptable tools embedded in the evolving dynamics of their broader environment. For instance, we show that while alliance partners most often define technological terms—such as product- and patent-related terms—in their contracts, stock-related definitions have become more common over the years.

2 Related work

From a linguistic standpoint, definitions formally ascribe meaning to an undefined term (i.e., the definiendum) using already established terms (i.e., definiens). An example illustrating this terminology is shown in Figure 1. By defining a term that is yet to be defined, definitions explicitly create a shared understanding among members who agree on the definition. In alliances, definitions set up in their contracts thus establish the meaning of terms relevant to the signing parties.

Given the importance of contract text and the definitions therein, existing AI research on automatically examining definitions is limited, primarily focusing on isolated definitions or specific terms. While this has generated a wealth of insights, as, for instance, Legg and Hutter in 2007, collect and compare 70 definitions of the word 'Intelligence' (Legg et al., 2007), it is limited because definitions are often embedded in real-world texts, such as alliance contracts.

Extracting information embedded within the broader contract text and accounting for variations in document layouts across real-world texts is challenging. While *Sentence Boundary Detection (SBD)* on curated data achieved accuracy scores of up to 99% as early as 2012 (Read et al., 2012), attaining such accuracy of SBD in noisy text data extracted from PDFs is an underexplored problem. Recent advances in the field are made by a series of shared tasks called FinSBD (Azzi et al., 2019; Au et al., 2020, 2021). The most successful method utilized two neural architectures, BiLSTM-CRF and BERT, and took into account both visual cues and textual data, achieving a mean F1-score of 0.937 on sentences and 0.844 on lists (Singh, 2020). Although this method can be taken as a proof of concept for extracting text from noisy PDFs, this approach does not scale for thousands of contracts. In turn, most work in the field focuses on specific information, which is easier to extract, such as the alliance partners of a contract (Sivapiran et al., 2023; Chalkidis et al., 2017). Extracting specific terms and meanings from contracts still largely depends on expensive expert annotation usually restricted to a subset of terms; for example, Hendrycks et al. (2021) annotated a subset of 41 categories.

3 Review of methods

Surveys of text representation techniques indicate that transformer-based embeddings currently perform best to compare semantic aspects between texts at scale in an unsupervised manner (Incitti et al., 2023; Patil et al., 2023). Thus, we selected two well-studied transformer-based sentence encoding methods. First, we relied on *Sentence-BERT (SBERT)* (Reimers and Gurevych, 2019), the most established model in prior literature. More specifically, we used the pretrained model *all_mpnet_base_v2*, which has achieved the highest overall quality in benchmarks[2] among SBERT models. As a more recent model, we choose the *General Text Embeddings (GTE)* Model (Li et al.,

[2] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

2023), as its largest variation *gte-large* was one of the best performing, freely available models on the Massive Text Embedding Benchmark (Muennighoff et al., 2023). We follow recent approaches of using clustered word embeddings as topic models (Sia et al., 2020). We utilize the BERTopic Framework (Grootendorst, 2022), which uses class-based TF-IDF to identify meaningful topics in the clusters. To ensure robustness of the topics, we use multiple Dimensionality Reduction Algorithms, mainly PCA (Lloyd, 1982) and UMAP (McInnes et al., 2018), as well as Clustering Methods (K-Means (Lloyd, 1982), BIRCH (Zhang et al., 1996), HDBSCAN (McInnes et al., 2017)) in combination.

4 Contract corpus and definition extraction

We collected $18,742$ alliance contracts in the biopharmaceutical industry signed between $1973 - 2021$ in PDF format retrieved from BioScience Advisors (now part of Evaluate), which consolidates information from the SEC and the Freedom of Information Act. This is structurally equivalent to the widely used Recap database (now Cortellis Deals Intelligence; (Hanisch et al., 2025; Schilling, 2009)). Before starting with our analysis, we conducted extensive manual cleaning efforts to ensure accuracy: removing duplicates, supplementing missing contracts via manual searches through Law Insider and the SEC, coding ancillary firm-level data, including headquarters and founding year from corporate websites, standardizing firm names, and accounting for name changes and parent-subsidiary relationships.

After collecting the contract texts in the form of PDF documents, we perform three steps to extract the definitions. First, we use the open source PyPDF2 Python package (Fenniak et al., 2022) to convert PDF documents to text. Generally, such tools generate noisy, unstructured texts containing nonstandard words, false starts, missing punctuation, missing letter case information, and other text disfluencies (Azzi et al., 2019), ultimately complicating the extraction of correct text passages. Therefore, we perform cleaning steps with regex to standardize whitespaces or tabs to one space length and remove multiple newlines in a row.

Second, we use three different methods to extract definitions. The first method to extract is based on multiple open source SBD tools, combined with a simple check for common words alliance partners use in definitions. The second method is based on a regular expression (regex) that searches for common words, common structure, and phrases of definitions. The third method is based on the Large Language Model `Llama-3.1-70B-Instruct` of the LLama 3 model family (Dubey et al., 2024). The regex and Large Language Model method are explained in more detail in Appendix Section A.

Third, we use a subset annotated by human coders to evaluate and compare the three extraction methods. Three independent coders, one author and two thesis students, all equipped with knowledge of the seminal alliance contract literature and information retrieved from interviews with two practitioners (more specifically, one alliance manager and one lawyer), identified and counted definitions in 826 alliance contract PDFs without predefined coding guidelines to maintain an open-ended qualitative approach. In rare cases of disagreement among the coders about the number of definitions, this was resolved through mutual discussion after completion of their individual coding efforts. In this subset, we have both contracts with zero definitions and larger contracts containing up to 735 definitions. Because of this large variance and contracts that are hard to extract data from due to the noisy PDF reading, we evaluate the methods by their median count error per contract against the human counts.

The SBD approach finds fewer definitions than the human-annotated dataset and has a median error rate of 6 definitions per contract. Additionally, this method fails to identify definitions in full length when the definition text spanned multiple sentences. In comparison, the regex method finds slightly more definitions than the human count and has a median error rate of 3 definitions per contract. Although the LLM method proved effective for shorter contracts, longer contracts introduced significant drawbacks. Even when setting the temperature to 0, the LLM still hallucinates, i.e., includes definitions that are not written in the real-world contract text. Similar problems are known for extracting facts from documents (Dong et al., 2022), where pretraining data is returned instead of document data and even larger models struggle to accurately return document facts, once the context size increases significantly (Li et al., 2024). Overall, the LLM method has a median error rate of 5 definitions per contract. Based on this comparison, we conclude to use the regex method for extracting

definitions from the whole corpus.

Using regex, we identify a total of 457,711 definitions across 12,468 contracts. In so doing, we identify 122,414 different definienda in the contract corpus. For the following analysis, we drop 77,482 definitions containing the phrase 'set forth in', which indicates that the definition is defined in another chapter of the contract. This leaves a final sample of 380,131 definitions. We release these definitions as a dataset[3] containing a unique id for each alliance contract, the year the contract was signed, and the definition itself, which is split into the definiendum and the definiens. For the published dataset, we use a regex to remove all company names from the text and replace them with the placeholder [COMPANY]. For the following analysis, these definitions were then preprocessed by removing all line breaks, standardizing whitespaces, and removing all non-alphanumeric characters.

5 Similarity of definitions

After identifying regular expressions as the most accurate method for extracting definition text from contracts, we aim to analyze the similarity of definitions. Specifically, we examine definitions with the same definiendum to understand which definitions are written in a standardized manner and which vary significantly. To do so, we select all definienda that appear in at least 100 different contracts (when converted to lower case), resulting in 369 definienda. After removing four redacted definienda (usually for confidentiality reasons), we retain 365 different definienda for further analysis. For example, the most usual definiendum, 'affiliate', is frequently defined (precisely in 6,584 different contracts).

Next, we create the embeddings solely on the definientia. Then, we calculate the mean of all pairwise cosine similarity scores between all definientia that have the same definiendum. We thus obtain a mean similarity score for each definiendum, where a high score indicates that definitions of a certain definiendum are defined more similarly across contracts, and a low score indicates the reverse.

In the following, we compared similarity across all definienda by relying on three inductively produced overarching categories of definitions (Gioia et al., 2013). In so doing, in a first step, we manually reviewed each contract definition individually and assigned first-order concepts, i.e., descriptive codes closely reflecting the wording in the data. After reviewing 74 contracts in this way, we aggregated these into second-order concepts by clustering related first-order concepts and abstracting their underlying themes. Finally, we distilled these second-order concepts into three mutually exclusive aggregate categories – social, technological, and organizational definitions. Table 1 displays similarity across definienda according to the GTE and SBERT embeddings.

GTE		SBERT	
Mean	Definiendum	Mean	Definiendum
0.961641	irs	0.909868	governmental order
0.960714	governmental order	0.901723	irs
0.960319	exchange act	0.882542	ema
0.957939	securities act	0.880703	exchange act
0.957101	ema	0.879796	erisa
...	...	...	...
0.883866	diligent efforts	0.603111	diligent efforts
0.883659	common stock	0.603016	technical information
0.883542	transaction documents	0.601633	research plan
...	...	...	...
0.821110	field of use	0.375891	field of use
0.818951	party	0.337282	party
0.816680	parties	0.320123	parties

Table 1: Similarity of definitions appearing in at least 100 different contracts according to GTE and SBERT embeddings.

The most similarly defined definienda are legal terms like 'irs' (i.e., internal revenue service) or 'ema' (i.e., European Medicines Agency). On the other hand, the most dissimilar definienda are specific to an alliance, e.g., the parties involved in the contract. For qualitative analysis, definienda which differ slightly across contracts are interesting, e.g. 'diligent efforts', which we analyze in Section 5.2.

Although individual score differences between the two embedding methods are substantial, the methods exhibit a strong overall rank correlation. The Spearman rank correlation between the two methods is approximately 0.96, and the Kendall tau correlation is approximately 0.84. The higher scores of GTE might indicate that GTE embeds the sentences in a similar embedding space, as they are all definitions, which SBERT might not do. However, as we only compare definitions with each other, this has little influence on the general ranking and further results are similar. For this reason, examples in this paper will only be shown when computed on GTE embeddings.

[3] https://huggingface.co/datasets/Maxbenkre/pharmaceutical_definitions

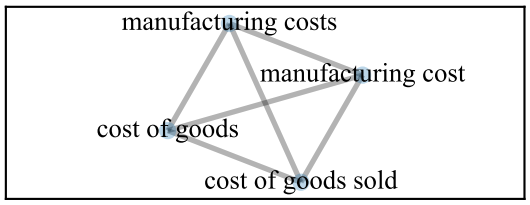

Figure 2: Similarity graphs: Connected definienda are defined with more than 99% cosine similarity according to GTE.

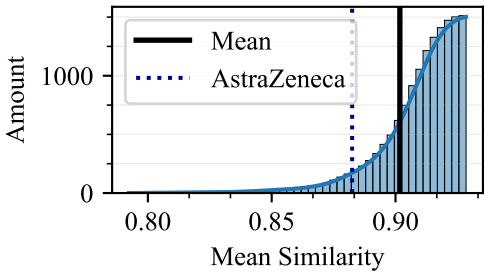

Figure 3: Cumulative distribution of mean pairwise similarity scores with GTE embeddings of various 'efforts' definitions compared to the AstraZeneca - EC definition.

5.1 Similarity graphs

Beyond comparing individual definitions, our objective is to identify different definienda that are described by textually similar definientia. For example, the terms 'cost of goods' and 'manufacturing costs' are often defined in a nearly identical manner across contracts. Successfully identifying such pairs validates our embedding model's ability to detect similarity.

Our method is to first compute a single embedding for each definiendum by averaging the embeddings of all its corresponding definientia. We then calculate the pairwise cosine similarity between these embeddings. A high similarity score between two embeddings indicates that their underlying definienda are used synonymously or at least described in a similar way.

We represent these relationships as a graph, where each term is a node and an edge connects two nodes if their similarity exceeds a set threshold. For GTE embeddings, a strict threshold of >0.99 identifies 193 distinct pairs of synonymous definienda. As shown in Figure 2, this reveals distinct clusters of meaning. The graph's structure is dynamic; for instance, lowering the threshold to 0.98 adds a new edge between the 'manufacturing cost' node and the related term 'development costs'. By incrementally lowering the threshold, we can explore relationships from almost equally defined definienda to more broadly related concepts. These similarity graphs (see more in Appendix B) can help new alliance partners to quickly find conceptually related definienda that use different terminology in their previous contracts, which can reduce unnecessary equivocality in further contracts.

5.2 Comparing similarity based on the example of 'reasonable effort' definitions

To demonstrate how semantic similarity can identify atypical legal definitions, we perform a case study on the term 'best reasonable efforts' from the contentious AstraZeneca-EC contract. As this term itself already is rare in our dataset—it is used in only three contracts—we benchmark its definition against a larger set of 1,512 definitions for the more common, related terms 'commercially reasonable efforts' and 'diligent efforts'. For this analysis, we generate GTE embeddings from the definiens only, excluding any redacted definitions.

First, we establish a baseline by calculating the mean pairwise cosine similarity across this entire 1,512-definition set, which is approximately 0.901. This high score indicates that the definitions for these standard "efforts" clauses are, on average, highly similar to one another.

We then calculate the mean pairwise similarity of the AstraZeneca definition against all 1,512 definitions in the comparison set. As shown in Figure 3, the resulting similarity score is 0.882. This value is significantly lower than the mean, confirming that the AstraZeneca definition is an outlier. Specifically, 1,346 of the 1,512 definitions in the comparison set have a higher mean similarity to their peers, placing the AstraZeneca definition in the 11th percentile for textual similarity and marking it as highly unusual.

We see that the 'best reasonable efforts' definition is infrequently used and highly dissimilarly defined compared to the other definitions. We manually analyze the differences for the reasons between the higher and lower scores. The definientia that are most similar to the majority of others are exemplary to the following one, with recurrent phrases highlighted:

'means, with respect to a Party's obligations under this Agreement, efforts *consistent with the efforts and resources normally used by a similarly sit-*

uated pharmaceutical, biotechnology or technology company in the exercise of its reasonable business discretion relating to the development or commercialization of a *product with similar product characteristics that is of similar market potential at a similar stage of development or commercialization* [...]'

So, the most common way to define 'efforts' is to compare them to efforts spent by similar companies that undertake similar projects. Usually, these definitions are very verbose; the 200 most similar definitions have a mean character length of 998. This seems to be one of the main differences when comparing definitions that have a lower similarity score.

The lowest similarity scores below 0.85 have definitions that are actually defined in a different section of the document, such as "shall have the meaning given in Section 2.2 (c)". When disregarding such definitions and comparing all definitions that have more than 0.86 similarity score but less than 0.89 score, it shows that these definitions are much less verbose. They contain only 637 characters on average instead. One such example of a particularly short definition is the following one, with a mean similarity score of 0.867: 'shall mean efforts in accordance with the standards of care and diligence [COMPANY] practices with respect to its own product'.

The definition in the AstraZeneca - EC contract, on the other hand, is as verbose as other similar definitions, as it contains 962 characters and includes the most common phrases. Instead, the low score could result from the inclusion of the urgent need for a vaccine for the pandemic:

'means (a) in the case of AstraZeneca, the activities and degree of effort that a company of similar size with a similarly-sized infrastructure and similar resources as AstraZeneca would undertake or use in the development and manufacture of a Vaccine at the relevant stage of development or commercialization having regard to the *urgent need for a Vaccine to end a global pandemic which is resulting in serious public health issues, restrictions on personal freedoms and economic impact, across the world* but taking into account efficacy and safety [...].'

This example demonstrates that examining the (dis)similarity of definitions enables alliance partners to distinguish between standard and atypical terms, whereas identifying the latter helps to proactively address potential misunderstandings early.

6 Change of definitions over time

Next, we investigate how contract definitions change over time. We first assess if there are definitions that have become more prevalent in general or specific terms (definienda) whose meaning (definiens) has changed over time. Second, we create topic models to identify prevalent topics and their trends.

6.1 Change of specific terms

We begin by examining how the use and meaning ascribed to specific terms evolve over time. This requires analyzing the frequency of definitions in different years. Figure 4 shows three key metrics from 1981 to 2021: the number of contracts containing definitions, the total number of definitions, and the average number of definitions per contract.

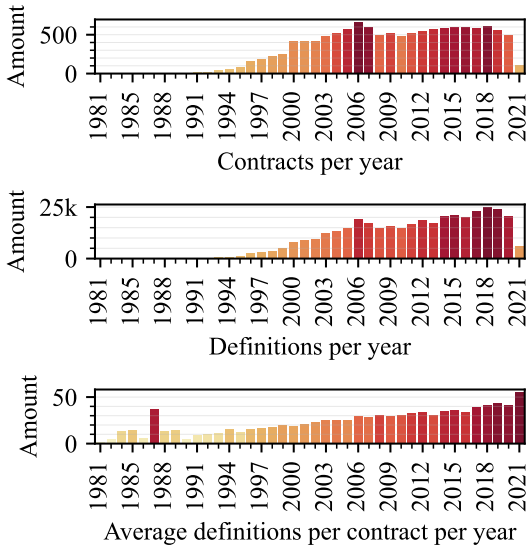

Figure 4: The top diagram shows how many contracts that contained definitions were found for each year. The diagram in the middle shows how many definitions were found for each year. The lower diagram shows the average number of definitions in each contract per year.

The upper part of the figure shows that the vast majority of contracts containing definitions in the dataset were signed from 2000 onward, with the oldest contracts signed in 1981 and the most recent ones signed in 2021. We limited our analysis to the period from 2000 to 2021, due to a rare number of definitions prior to the year 2000. The middle section of the figure illustrates the annual number of definitions included in the contracts. Comparing this with the upper section of the figure reveals an interesting trend: while the majority of con-

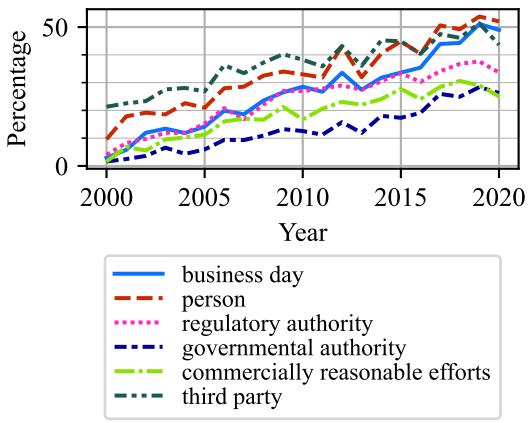

Figure 5: Definienda with the highest annual percentage increase in the proportion of contracts they appeared in.

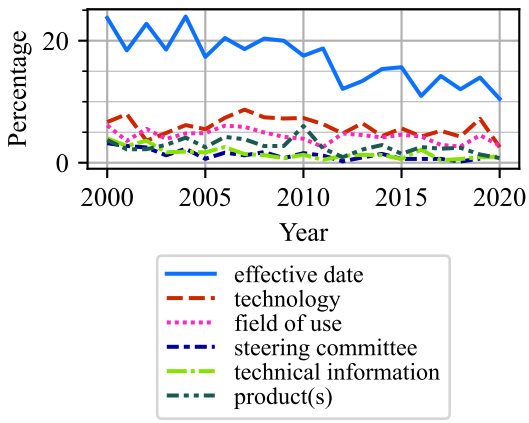

Figure 6: Definienda with the highest annual percentage decrease in the proportion of contracts they appeared in.

tracts containing definitions were signed in 2006 and 2008, the highest number of definitions appears in contracts signed between 2016 and 2019. This suggests that alliance partners have increasingly included more definitions per contract over time while the variation has potentially increased as well. Specifically, although the number of contracts containing definitions peaked in 2006 and 2008 (i.e., it was usual to include definitions in that time span), the average number of definitions per contract increased in subsequent years. Interestingly, the peak in contracts that contain definitions coincides with the financial crisis, highlighting a potential link between the macroeconomic environment and contracting.

Starting with a more fine-grained analysis, we first uncover temporal patterns of definienda in terms of their frequency, i.e., whether and how often definienda reoccur over time. As the absolute number of definitions has increased, as shown in Figure 4, the measure must take into account the number of contracts in each year. Figure 5 shows the definienda that experienced the most significant relative increase compared to all other definitions appearing in at least 100 different contracts between 2000 and 2020. In total, 35 definitions show an increase of 10 percentage points or more during this period. The definition with the highest increase is ‘business day’, which was defined in only 3% of the contracts signed in 2000 but appeared in nearly 50% of contracts signed in 2020. Other definienda with substantial increases above 20% include ‘person’, ‘commercially reasonable efforts’, ‘third party’ and two different authority-related terms.

As shown in Figure 6, very few definitions are used significantly less frequently between 2000 and 2020. The only definiendum that decreased by 10 percentage points in frequency during this period is ‘effective date’. Other definitions show, at most, a decline of five percentage points. However, there are notable outliers, such as the definiendum ‘emea’, which surged from 2% occurrence in 2000 to 15% in 2009, and then plummeted to 0% by 2019. In comparison, ‘ema’ experienced the highest single-year increase, rising from 0% in 2009 to 9.6% in 2009. The decline in ‘emea’ and the parallel increase of ‘ema’ can be attributed to the renaming of the EU agency responsible for the scientific evaluation, supervision and safety monitoring of medicines from EMEA (European Medicines Evaluation Agency) to EMA (European Medicines Agency) in 2010.

To analyze changes in definitions over time, we applied cosine similarity to compare embeddings of definitions that appeared in at least 30 different contracts between 2000 and 2010 and compared them to embeddings of definitions from 2011 to 2020. For each definiendum, we calculate the mean embedding for each decade and then computed the cosine similarity between the two mean embeddings. However, this method has a limitation: definitions that are inherently diverse, such as those for ‘company’, which are often specific to individual alliances, are likely to exhibit low cosine similarity between decades regardless of actual semantic change. Therefore, the results must be interpreted cautiously to distinguish actual changes in definitions from definitions being diverse in the first place.

However, between the decades, there are only three definitions that are outliers and have a significantly lower cosine similarity compared to the distribution of other definitions, indicating that few definitions had major shifts in the way they were defined (i.e., the meaning ascribed to the same terms) over time. These definienda were 'holder', 'active ingredient', and 'cmc'. In the cases of 'holder' and 'cmc', the definitions in newer contracts were standardized, i.e., experienced a convergence. Before that convergence, the definition experiences changes. For example, the definientia of 'holder' in the years before 2010 (i.e., before the convergence) are 'means any Investor who holds at least 200,000 shares of Preferred Stock [...]' or 'shall mean a Person holding Company Common Stock [...]' and examples after 2010 are 'means a Person in whose name a CVR is registered in the CVR Register' and 'means a Person who is registered in the CVR Register'. The term 'CVR' in these cases usually is defined as the right of holders to receive contingent Parent Common Stock. The definientia for 'cmc' after 2010 were shortened to 'means chemistry, manufacturing and controls' from 'means the Chemistry Manufacturing and Controls [...] as filed with the FDA'. In the case of 'active ingredient' the major shift was that the actual substance used was rarely defined in the years after 2010 and often defined in the years before. This means that, concerning the technology, alliance partners decided to include less actual detail. Two examples from before 2010 are 'shall mean Mesalamine [...]' and 'shall mean the doxycycline hyclate [...]'. Two examples from after 2010 are 'means the clinically active material(s) that provide pharmacological activity in a pharmaceutical product [...]' and 'means any substance (whether chemical or biologic) or mixture of substances intended to be used in the manufacture of a drug [...]'.

These examples underscore that this method finds interesting changes in definitions. However, it does not guarantee finding all definitions that have changed significantly over the years.

6.2 Dynamic topic models

Building on the observation that definitions have generally become more common overall, we next investigate whether this trend is the same across different categories of definitions, or if a certain category has become more prevalent over the years. To do this, we apply dynamic topic modeling using BERTopic (Grootendorst, 2022) to all definitions

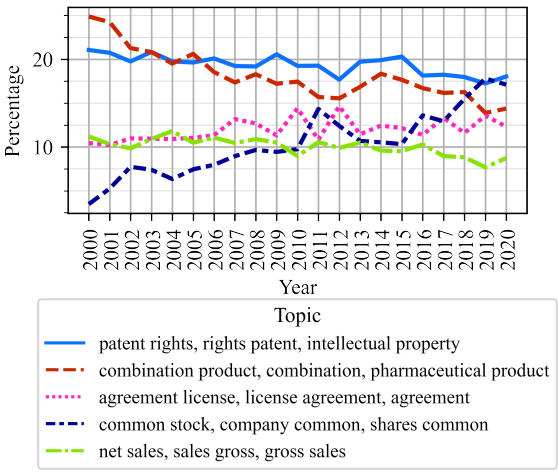

Figure 7: Top five topics of formation contract definitions of a topic model created with UMAP, K-Means and GTE embeddings.

from biopharmaceutical alliance contracts signed between 2000 and 2020. Model performance in regards to cluster size is evaluated using the standard approach of topic coherence and topic diversity (Wu et al., 2024), and we use multiple different approaches, as detailed in Appendix C.

The analysis revealed a dominant trend as displayed in Figure 7: definitions related to 'stock' and 'shares' became significantly more prevalent over time. This rise in financial terminology, likely driven by increased M&A activity (DiMasi, 2020), coincided with a decline in the frequency of definitions related to technological product development.

7 Conclusion

Overall, this paper applies NLP methods to a unique corpus, whose central role in forming alliances makes them highly relevant for both practice and research. Specifically, we extracted definitions from real-world alliance contract texts in the biopharmaceutical industry at scale. First, methodologically, we develop a tool set that allows us to study definitions in text corpora using a combination of traditional tools and modern machine learning approaches. More specifically, we find that regex seems most suitable for extracting definitions from noisy real-world contract text. Second, we contribute to the research community studying computational processing of language by publishing a comprehensive dataset of definitions, distinguishing between definienda and definientia in alliance contracts over time. Third, this interdisciplinary study also hopes to make two theoretical contribu-

tions to existing research. While our results show that the ascribed meaning to social terms are most dissimilar and social definienda are most unusual (i.e., they differ the most in meaning and usage), our method allows to derive whether a certain definition differs from the most usual way it is defined across alliance contracts. This, in turn, could provide alliance partners with the ability to identify unusual definitions or dissimilarly defined terms, allowing them to focus on those definitions in negotiations to navigate potential conflicts early on. In addition, we analyze definitions in alliance contracts signed between 2000 and 2021 and find that alliance partners generally use more definitions over time. Although the number of definitions used per contract increased over time, the content of definitions is embedded in the broader environmental context; for instance, stock-related definitions have become more common after the financial crisis. The latter extends more traditional views on alliance contracts, which have mainly focused on the transaction itself, by interpreting them as embedded in larger economic surroundings. In conclusion, in applying NLP methods to real-world corpora, our paper integrates qualitative insights with quantitative analysis. While we acknowledge that qualitative information has inherent limitations - for instance in terms of generalizability - our study postulates and aims to showcase that such human-centered qualitative analyses support, contextualize and embed NLP findings in a meaningful way, particularly in context-rich texts such as alliance contracts.

Limitations

Our paper has three limitations that potentially open avenues for future research. While the text extraction of definitions from real-world texts (i.e., alliance contracts) is a key strength of our paper, this extraction also comes with limitations. First, while the regex method worked reasonably well when compared to the dataset of human annotations, there is no guarantee that the extracted definitions are complete or correct. As the regex exploited the most common structure found in most contracts, definitions of contracts following different formats might not be found. As the text read from the PDF files already contains text anomalies, the definitions found by the regex will contain them as well and can negatively impact the similarity scores and topic creation.

Second, similarity is highly dependent on the viewpoint. While syntactic similarity, which has been at the heart of linguistic research, is rather objective, semantic similarity might vary. For example, a lawyer might interpret the similarity of two definitions differently than the CEO of an alliance partner or a stakeholder of an alliance partner. Our method to identify similarity should therefore not be taken at face value, but certainly provides a first step to filter interesting definitions for further in-depth research.

Third, we assume that definitions only concern one topic each. While this is a reasonable assumption, it might not hold true for all definitions. For example, patent definitions can be related to technological and legal topics. Future research could thus use other topic model approaches like latent dirichlet allocation or soft clustering methods to adjust for this.

References

Willy Au, Abderrahim Ait-Azzi, and Juyeon Kang. 2021. Finsbd-2021: The 3rd shared task on structure boundary detection in unstructured text in the financial domain. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 276–279, New York, NY, USA. Association for Computing Machinery.

Willy Au, Bianca Chong, Abderrahim Ait Azzi, and Dialekti Valsamou-Stanislawski. 2020. FinSBD-2020: The 2nd shared task on sentence boundary detection in unstructured text in the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 47–54, Kyoto, Japan. -.

Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, Macao, China.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28, New York, NY, USA. Association for Computing Machinery.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Joseph A DiMasi. 2020. Research and development costs of new drugs. *JAMA*, 324(5):517–517.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and PyPDF2 Contributors. 2022. The PyPDF2 library.

Dennis A. Gioia, Kevin G. Corley, and Aimee L. Hamilton. 2013. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. *Organizational Research Methods*, 16(1):15–31.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Marvin Hanisch, Lorenz Graf-Vlachy, Carolin Haeussler, Andreas König, and Theresa S Cho. 2025. Kindred spirits: Cognitive frame similarity and good faith provisions in strategic alliance contracts. *Strategic Management Journal*, 46(2):436–469.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Francesca Incitti, Federico Urli, and Lauro Snidaro. 2023. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Shane Legg, Marcus Hutter, et al. 2007. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Ian R Macneil. 1978. Essays on the nature of contract. *NC Cent. LJ*, 10:159.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*, PP:1–1.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Melissa A Schilling. 2009. Understanding the alliance data. *Strategic Management Journal*, 30(3):233–260.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Janvijay Singh. 2020. PublishInCovid19 at the FinSBD-2 task: Sentence and list extraction in noisy PDF text using a hybrid deep learning and rule-based approach. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 55–61, Kyoto, Japan. -.

Sanjeepan Sivapiran, Charangan Vasantharajan, and Uthayasanker Thayasivam. 2023. Party extraction from legal contract using contextualized span representations of parties. In *Proceedings of the 14th*

International Conference on Recent Advances in Natural Language Processing, pages 1085–1094, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: an efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114.

A Extraction methods

We explain the regular expression and Large Language Model approach.

A.1 Regular Expression

The final regex we use to extract definitions is displayed in Figure 8. We explain the purpose of each part of the regex in the following:

- ("|“)((.|\n){1,100}?)("|”): This part matches the definiendum. It works by first matching the opening quotation mark, which can have different formats. Then up to hundred characters can follow and the quotation mark has to be closed again.
- ((.|\n){0,50}?)(mean|define): This part matches everything up to the connector phrase. The lazy quantifier is necessary as otherwise this could match the connector phrase of the next definition and two definitions would be extracted as one.
- ((.|\n)*?)\.: This part matches everything that comes after the connector phrase until a terminal point occurs. Note that this group combined with the previous one can also match different forms of connectors, e.g ‘has the meaning’ or ‘is defined as’, as all characters before and after the character sequence ‘mean|define’ are matched.
- (\n|\s\s+): This part is used to see if the point that was found is indeed a terminal point finishing the sentence. This is necessary as it is difficult to distinguish abbreviations from terminal points otherwise. On top of that, it is a way to match both single sentence definitions and multiple sentence definitions as long as there is only one white space character between the sentences.

Here are two definitions from different contracts and how they are detected by different parts of the regex:

- 1.23 “Product” shall mean any product containing a Development Candidate.¬
- “API Manufacturing Process” is defined as a process used in the manufacture of API [. . .] in order to market, sell and distribute the Product in the Territory.¬

We display the comparison of human count and regular expression count in the upper part of Figure 9. We do an error analysis of 26 contracts with a difference of fifty or more from the manual count. We find that thirteen contracts do not use quotation marks, four contracts had no connector phrases e.g. “mean” or “define”, two contracts exhibited poor OCR layer quality, and one contract featured multiple definitions defined more than once. Additionally, in six cases, we find that the manual count was inaccurate, showing that the different layouts between contracts and the repetitive nature of the task can even make humans inaccurate.

A.2 Large Language Model

We employ `Llama-3.1-70B-Instruct` (Dubey et al., 2024) to extract definitions from legal contracts. Consistent with previous research (Li et al., 2024; Dong et al., 2022), every prompt variation we test generates definitions that are not present in the source text, particularly for longer contracts.

Several prompting techniques fail to mitigate this issue. In-context learning proves counterproductive, causing the model to invent definitions that are similar to our provided examples, even if they are not present in the contract. Structuring the output into a separate definiendum and definiens resulted in pairs, where the definiendum was part of the original contract, but the definiens was not. Furthermore, chunking the contract into smaller segments to reduce the input context led to an even higher rate of false definitions, particularly in sections of the contract which contained no definitions at all.

While the prompt detailed in Figure 10 achieves the best performance relative to the human baseline, it still produces hallucinated definitions. As hallucinated definitions arguably are more detrimental for further research, compared to missed definitions and other research reports similar results (Li et al.,

```
("|“)((.|\n){1,100}?)("|”)((.|\n){0,50}?)(mean|define)((.|\n)*?)\.(\n|\s\s+)
```

Figure 8: The full regular expression we use to extract definitions from contracts.

2024; Dong et al., 2022) we do not pursue LLMs further for this task.

We display the comparison of human count and the number of definitions found by the LLM in the lower part of Figure 9. Similar to our analysis on a subset of contracts we see that the LLM returns more definitions than the human annotators counted. We analyze a subset of contracts and find that for many contracts the LLM hallucinates definitions that are not part of the actual contract text.

B Similarity graphs

We average the embeddings for all definientia of one definiendum and then calculate the pairwise cosine similarity between different definienda. Depending on the method, different thresholds can be set to judge if a pair of definienda are similar to each other. For example, setting the threshold of similarity above 0.99 for GTE embeddings finds 193 pairs of similar definienda. By selecting each definiendum as a node in a graph and connecting nodes when their definienda are deemed similar, it is possible to easily visualize which definienda are similarly defined according to this method.

To show that the found pairs are similar to each other we show a subset of graphs here. There are 72 distinct graphs that can be created with this method, when setting the threshold above 0.99 cosine similarity for GTE embeddings. In Figure 13 we show all 21 graphs that have more than 2 nodes and fewer than 9 nodes. There are 2 graphs with more than 8 nodes and 49 graphs with only two nodes not depicted.

It is important to note, however, that definienda that are defined similarly are not necessarily semantically equivalent. One example are the definienda ‘excluded assets’ and ‘purchased assets’ which actually describe opposite terms. The definientia of both terms are syntactically very similar however, e.g. the following one for excluded assets: *“shall mean all assets and properties (other than the Purchased Assets)”* is very similarly defined to the one for the purchased assets of a different contract: *“shall mean all assets of Seller (excluding only the Excluded Assets described in Section 2.2 below)”*. In this case the embedding rated the similarity highly, even though the semantic meaning is actually opposite.

C Topic model evaluation

Topic model effectiveness is typically evaluated by using two metrics: topic coherence and topic diversity. A widely accepted measure for topic coherence is *Normalized Pointwise Mutual Information (NPMI)* (Bouma, 2009), which has demonstrated reasonable performance in mimicking human judgment (Lau et al., 2014). The measure ranges from $[-1, 1]$, where 1 indicates a perfect association. Topic diversity is the percentage of unique words for all topics (Dieng et al., 2020). This metric is quantified within the range $[0, 1]$, where a value of 0 indicates completely redundant topics and a value of 1 indicates completely diverse topics. It is important to acknowledge that, while these measures provide an indication of a model’s performance, they serve as proxies for what is ultimately a subjective evaluation. Coherence and diversity can be perceived differently depending on the individual, and these measures should not be used as a way of optimizing topic models but instead, they should be viewed as providing an indication of performance (Grootendorst, 2022).

Both NPMI and diversity scores are calculated for each cluster amount, ranging from 10 to 50 clusters. As HDBSCAN cannot create a specific amount of cluster, the minimum cluster size hyperparameter is set between 300 and 1000 in steps of 17 so that between 10 and 50 clusters are created. The dimension to which the dimensionality reduction algorithm reduces to is tested for each algorithm in steps of 10 from 10 to 50. Other hyperparameters are kept constant and to their default values during different runs. As coherence score calculation is memory intensive, 100,000 definitions are randomly chosen as a subset on which all clusters are created.

Another aspect of clustering is the running time of each model, as faster models are more practical to use. For this matter, all combinations of Dimensionality Reduction and Clustering Methods are used with hyperparameters set to 10 clusters

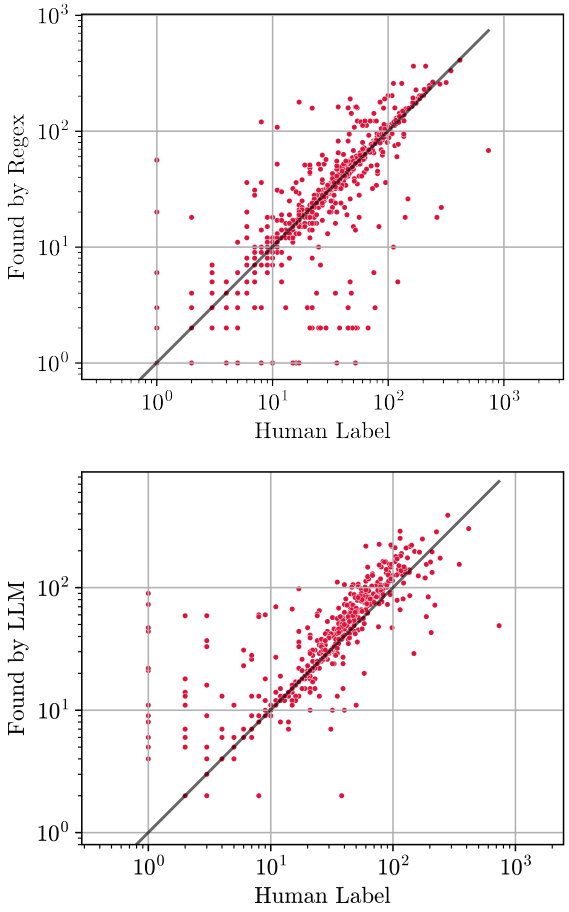

Figure 9: Comparison of human count to the number of definitions found by regex in the upper figure and by LLM in the lower figure on log scale. The black line indicates perfect alignment, while the each dot above the line means the corresponding method found more definitions than the human count, while each dot below means the humans counted more definitions than the method found. We can see that in the majority of contracts the LLM finds more definitions than the human count. In our analysis we find a lot of hallucinated definitions that are not part of the actual contract.

or 1000 as the minimum cluster size for HDBSCAN. The time we need to embed the definitions was not added, as this step can be precomputed. This one time step takes around 5 GPU hours for each embedding method on the whole dataset. All of these timed runs are performed on an `NVIDIA GeForce RTX 2080 Ti` and an `Intel(R) Core(TM) i7-9700K CPU` and we reporte the average over 3 runs.

The results of the evaluation can be seen in Table 2. The HDBSCAN clustering algorithm performs the best out of the three methods tested, both in terms of general topic coherence and diversity, while BIRCH performs worst. However, the results

LLM Prompt

```
System: Your task is to find all definitions in contracts. Respond only in python list format. Add a new entry to the list for each definition. Highlight the definiendum in each definition with «». If you don't find any return an empty python list.
[
"«Definiendum1» means ...", // The first definition
"«Definiendum2» means ...", // The second definition
]
User: <Text of one contract>"
```

Figure 10: The prompt with the best performance for extracting definitions with a Large Language Model.

of HDBSCAN should not necessarily be taken at face value, as it filters out noisy data compared to the other clustering algorithms tested. Depending on the number of clusters filtered by their size, HDBSCAN only clusters between $50 - 70\%$ of the definitions.

In general, the definitions are quite diverse, as all diversity scores are greater than 0.9. This is underlined by the coherence scores of different cluster sizes. On average, topic models with 30 clusters have a coherence score 0.07 higher than topic models with 10 clusters, and topic models with 50 clusters have an additional increase in the coherence score of 0.03.

The best coherence scores are achieved with a large number of clusters, and we show the top five topics of the topic models with 50 topics. We show the dynamic topic models for each type of clustering algorithm that we use with the fixed-dimensionality reduction method UMAP. Figure 11 shows the topics for HDBSCAN, Figure 7 shows the topics for K-Means, and Figure 12 shows the topics when using BIRCH. All topic models show the trend that technological definitions about products or their development were the most common in the year 2000, but became less common over the years. In contrast, stock-related definitions have become much more common compared to other definition topics.

Embedding	Dimens. Reduction	Clustering	Coherence	Diversity	Runtime
GTE	PCA	BIRCH	-0.019439	0.961011	25.55
GTE	PCA	HDBSCAN	0.083254*	0.983011	164.85
GTE	PCA	K-Means	0.003356	0.955560	30.80
GTE	Truncated SVD	BIRCH	-0.038055	0.962230	25.00
GTE	Truncated SVD	HDBSCAN	0.082802*	0.983106	159.85
GTE	Truncated SVD	K-Means	0.001938	0.955537	24.33
GTE	UMAP	BIRCH	0.020336	0.960023	116.29
GTE	UMAP	HDBSCAN	0.052988*	0.951642	119.28
GTE	UMAP	K-Means	0.028103	0.968145	116.53
SBERT	PCA	BIRCH	-0.014009	0.912952	70.84
SBERT	PCA	HDBSCAN	0.128869*	0.978100	166.38
SBERT	PCA	K-Means	0.015207	0.938769	23.03
SBERT	Truncated SVD	BIRCH	-0.020028	0.909377	70.05
SBERT	Truncated SVD	HDBSCAN	0.123977*	0.979926	164.95
SBERT	Truncated SVD	K-Means	0.014620	0.938340	22.34
SBERT	UMAP	BIRCH	0.021142	0.952062	115.29
SBERT	UMAP	HDBSCAN	0.096806*	0.959011	118.83
SBERT	UMAP	K-Means	0.039753	0.963235	113.36

Table 2: The mean coherence and diversity scores achieved by the combination of multiple methods and their corresponding average runtime in seconds. HDBSCAN coherence scores are marked as it only clusters between $50 - 70\%$ of definitions.

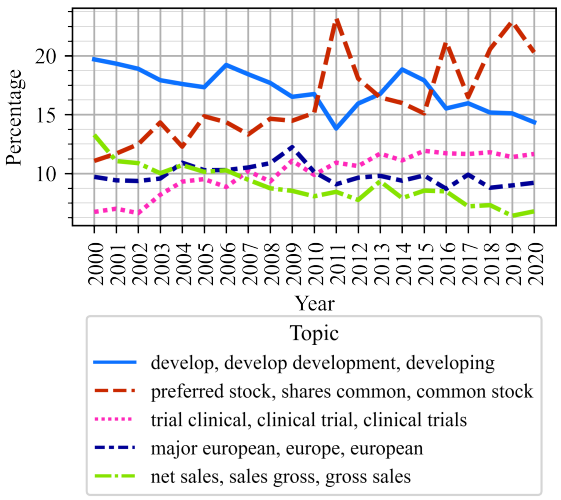

Figure 11: Top five topics of a Dynamic Topic Model using UMAP, HDBSCAN and GTE Embeddings.

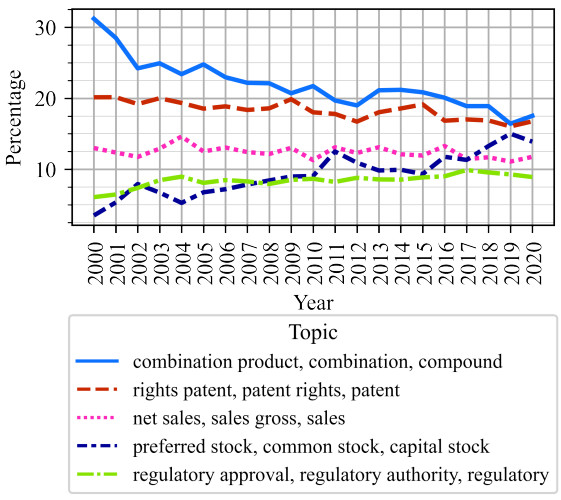

Figure 12: Top five topics of a Dynamic Topic Model using UMAP, BIRCH and GTE Embeddings.

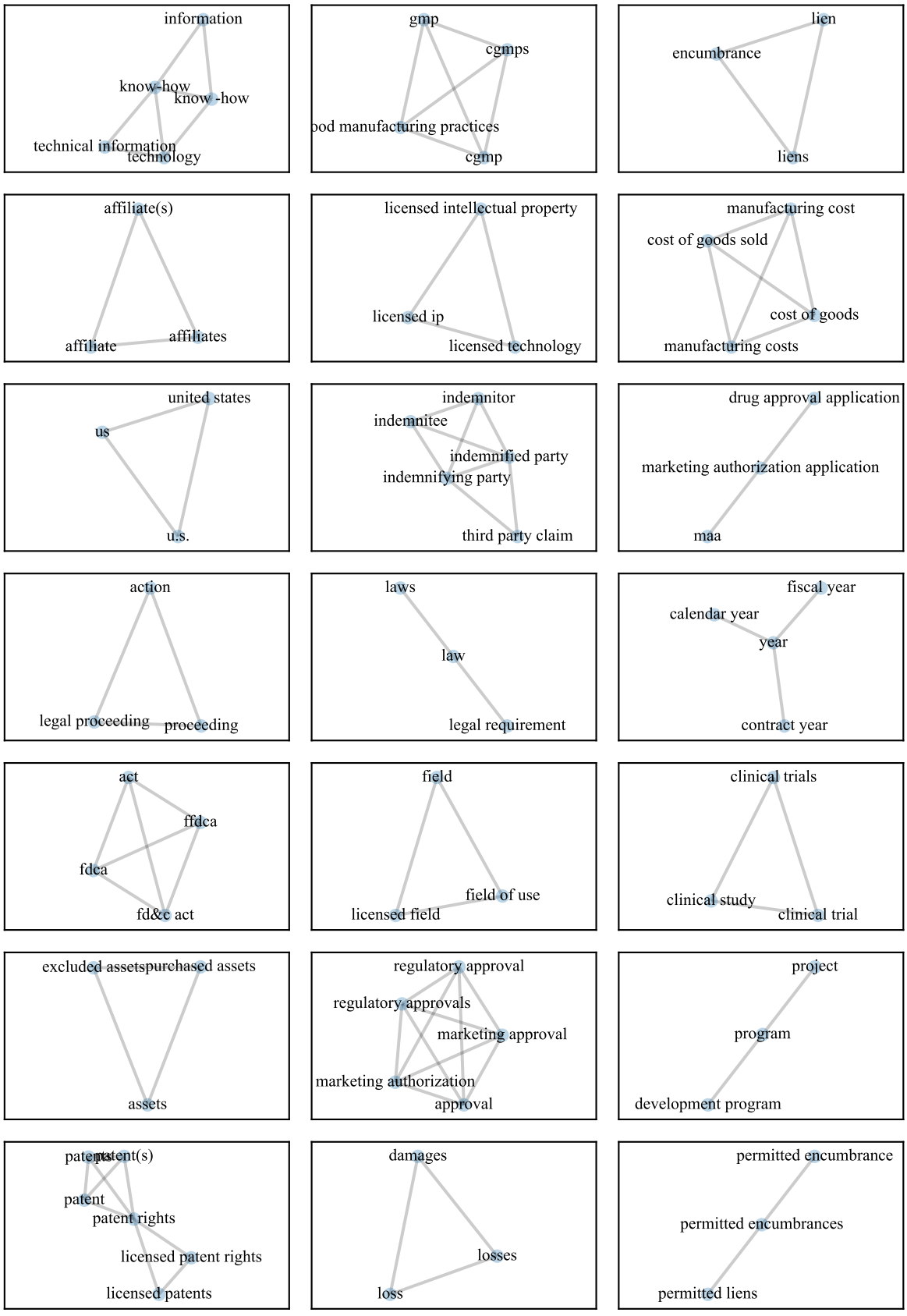

Figure 13: Similarity graphs: Connected definienda are defined with more than 99% cosine similarity according to GTE.