Towards Better Evaluation for Generated Patent Claims

Lekang Jiang[†], **Pascal A Scherz**[◇], **Stephan Goetz**[†]
[†]University of Cambridge, [◇]PSPB Patent Law
{lj408, smg84}@cam.ac.uk, post@pspb.eu

Abstract

Patent claims define the scope of protection and establish the legal boundaries of an invention. Drafting these claims is a complex and time-consuming process that usually requires the expertise of skilled patent attorneys, which can form a large access barrier for many small enterprises. To solve these challenges, researchers have investigated large language models (LLMs) for automating patent claim generation. However, existing studies highlight inconsistencies between automated evaluation metrics and human expert assessments. To bridge this gap, we introduce Patent-CE, the first comprehensive benchmark for evaluating patent claims. Patent-CE includes comparative claim evaluations annotated by patent experts, focusing on five key criteria: feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. Additionally, we propose PatClaimEval, a novel multi-dimensional evaluation method specifically designed for patent claims. Our experiments demonstrate that PatClaimEval achieves the highest correlation with human expert evaluations across all assessment criteria among all tested metrics. This research provides the groundwork for more accurate evaluations of automated patent claim generation systems.[1]

1 Introduction

The patent literature serves as a critical documentation of technological innovation (Mossoff, 2000). Patents are legal documents that grant exclusive rights to inventors in exchange for public disclosure of their inventions (Frumkin, 1947). The patent claims are the most legally significant section of a patent document, as they define the scope of protection and delineate the boundaries of an invention from known techniques to ensure legal enforceability (European Patent Office, 2000). Drafting precise and effective patent claims is a challenging task, which requires not only technical expertise but also an understanding of legal language and jurisdiction-specific regulations (Faber, 1990). Unlike general-purpose texts, patent claims must be both broad enough to encompass potential variations of an invention and specific enough to withstand legal scrutiny. This complexity often necessitates the involvement of skilled patent attorneys, which often renders the process both time-intensive and costly (LLP, 2023).

In response to these challenges, research has explored automated methods for patent claim generation to support inventors and attorneys. Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of general and patent-related tasks (Zhao et al., 2023; Jiang and Goetz, 2025). For instance, Jiang et al. (2025c) examined whether LLMs could generate high-quality patent claims based on patent descriptions, while another work investigated whether LLMs could revise patent claims to improve quality (Jiang et al., 2025b). These studies aim to accelerate the claim drafting process and reduce associated costs.

Despite these advances, a reliable automated evaluation for the quality of generated patent claims remains an unresolved challenge. Previous studies have relied on human expert evaluations, which are both time-consuming and costly, and they revealed inconsistencies between existing automated metrics and human assessments (Zuo et al., 2024; Jiang et al., 2025c,b). Table 1 highlights the fundamental differences between patent claim evaluation criteria and existing text evaluation methods. While current evaluations often rely on sequence overlap, semantic similarity, or multi-dimensional criteria such as coherence and fluency, patent claims have unique language requirements. Such requirements, such as the consistent use of terminology, technical formality, and high-level precision, are not common in other types of texts. Therefore, these differences underscore the limitations of existing

[1] https://github.com/scylj1/PatClaimEval

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3775–3788

Reference claims: 1. A mobile communications device comprising: a communication subsystem for communicating with a network; a microprocessor operably connected to the communication subsystem and to a memory; a local common address database accessible to a plurality of applications on the mobile communications device; and ... 2. The mobile communications device of claim 1 wherein ... ... **Candidate claims:** 1. A mobile communications device comprising: a communication subsystem configured to facilitate communication with a network; a processor operably connected to the communication subsystem and to a memory, the memory containing a local common address database and instructions that ... 2. The mobile communications device of claim 1, wherein ... ...
N-gram-based evaluations (measuring sequence overlaps): BLEU ROUGE METEOR ...
Embedding-based evaluations (measuring semantic similarities): BERTScore BARTScore MoverScore SimCSE ...
Multi-dimensional evaluations: UniEval (Coherence, Consistency, Fluency, Relevance) AlignScore (Factual consistency) ...
Human evaluations for patent claims: Feature completeness Conceptual clarity Terminology consistency Logical linkage Overall quality

Table 1: Comparison between current automatic text evaluation metrics and patent claim evaluation criteria. Patent claims have specific requirements different from other texts, which makes the evaluation difficult.

metrics and highlight a significant gap in the reliability and validity of automated evaluation methods for patent claims.

In this paper, we present the first benchmark for patent claim evaluation and propose a novel evaluation method tailored to the unique requirements of patent claims. The main contributions of this work are as follows:

1. We present Patent-CE, the first comprehensive benchmark for patent claim evaluation. Patent-CE includes 1,228 data points, which consist of a reference claim, two candidate claims, and comparative evaluations annotated by patent experts in five dimensions: feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality.

2. We propose a novel multi-dimensional evaluation method for patent claims, named PatClaimEval. PatClaimEval leverages Longformer (Beltagy et al., 2020) as its backbone and is trained on our dataset using a variation of contrastive learning (Gao et al., 2021).

3. We demonstrate the effectiveness of PatClaimEval through extensive experiments. Our results show that PatClaimEval achieves the highest correlation with human expert evaluations across all assessment criteria compared to existing metrics, including 6 N-gram-based methods, 4 embedding-based approaches, 2 multi-dimensional evaluators, and 1 LLM-as-a-judge method.

By tackling the evaluation problem, this research paves the way for more reliable assessments of automated patent claim generation systems, ultimately contributing to advancements in this emerging field.[2]

2 Related Works

2.1 Patent Claim Generation

Some studies have explored LLMs for automatically generating patent claims. An early investigation by Lee and Hsiang (2020) served as a preliminary effort and focused on fine-tuning GPT-2 (Radford et al., 2019) to generate patent-like texts. The authors found that minimal training was sufficient to produce patent-like outputs but did not assess the quality of the generated text. Building on this, Lee (2020) trained GPT-2 to transform one section of a patent application into another, such as the generation of abstracts from titles or claims from abstracts. However, since abstracts are often generic and imprecise, the generation of claims from abstracts may not be a well-conditioned task.

Therefore, Jiang et al. (2025c) introduced a description-based claim generation task and evaluated the performance of various LLMs on this

[2]Notably, we focus on reference-based claim evaluations, which is different from the real patent examination. We introduce details in the Limitation section.

Dataset	Task	Domain	Evaluation Criteria
QAGS (Wang et al., 2020)	Summarization	News	Factual consistency
SummEval (Fabbri et al., 2021)	Summarization	News	Fluency, coherence, consistency, relevance
Persona-Chat (Zhang et al., 2018)	Dialogue generation	General	Fluency, engagingness, consistency, personalization
Topical-Chat (Mehri and Eskenazi, 2020)	Dialogue generation	General	Naturalness, coherence, engagingness, groundedness, understandability
ToTTo (Parikh et al., 2020)	Table-to-text generation	General	Fluency, faithfulness, coverage
Patent-CE (Ours)	Patent claim generation	Patent	Feature completeness, conceptual clarity, terminology consistency, logical linkage, overall quality

Table 2: Comparison of commonly used benchmarks for text generation evaluation. Patent claims have unique language requirements different from other type of texts.

domain-specific challenge. Their human evaluation by patent professionals highlighted the limitations of various LLMs in generating high-quality patent claims and revealed inconsistencies between automated and human evaluation metrics. While previous studies tested the models on U.S. patents, Jiang et al. (2025a) further investigated the claim generation task on European patents. Additionally, Jiang et al. (2025b) extended the research to claim revision, investigating whether LLMs could further enhance claim quality.

Another study proposed the task of next-claim generation (Zuo et al., 2024), which involves generating the second and/or third claims based on the first claim. Likewise, this research also demonstrated a weak correlation between automated and human evaluation results for the next-claim generation task.

2.2 Benchmarks for Text Generation Evaluation

Accurately and efficiently evaluating the quality of generated texts is important for developing text-generation LLMs. Researchers have introduced several text evaluation datasets across various domains, each with distinct evaluation criteria. Table 2 summarizes some commonly used benchmarks for text generation evaluations.

Datasets, such as QAGS (Wang et al., 2020) and SummEval (Fabbri et al., 2021), are designed to evaluate summarization tasks within the news domain. These benchmarks focus on criteria including factual consistency, fluency, coherence, and relevance to ensure the generated summaries accurately represent the source text while maintaining readability. In addition, dialogue systems have benefited from benchmarks such as Persona-Chat (Zhang et al., 2018) and Topical-Chat (Mehri and Eskenazi, 2020), which target open-domain conversational tasks. Persona-Chat emphasizes personalization, fluency, and engagingness, while Topical-Chat introduces evaluation metrics for naturalness, coherence, and groundedness to advance the development of more realistic and context-aware conversational AI. Furthermore, the ToTTo dataset (Parikh et al., 2020) supports the task of converting structured tables into natural text. It evaluates fluency, faithfulness, and coverage to ensure the generated text aligns accurately with tabular inputs and effectively conveys the intended information.

Our Patent-CE dataset is specifically designed for the task of patent claim generation. Unlike other benchmarks, Patent-CE emphasizes feature completeness, clarity, terminology consistency, and logical linkage. All these critical aspects are for the legal and technical precision required in patent documentation. This dataset fills an essential gap by providing a benchmark tailored to the patent domain, presenting unique challenges not encountered in general-domain tasks.

3 Dataset

3.1 Human Annotation

Experienced patent experts were provided with reference and candidate patent claims to evaluate. Their evaluation was based on five aspects, adhering to established examination criteria (Jiang et al., 2025c): feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. These evaluation aspects are consistent with previous research and defined as follows.

(1) Feature Completeness: The extent to which the generated claims encapsulate all critical aspects of the invention. **(2) Conceptual Clarity**: The clarity and unambiguity of the language used in the claims. **(3) Terminology Consistency**: The uniformity in the use of terms throughout the claims. **(4) Correctness of Logical Linkages**: The accuracy with which features are interconnected and related. **(5) Overall Quality**: An aggregate measure that combines all the above criteria. Detailed evaluation instructions can be found in Appendix A.

3.2 Construction

To create a comprehensive dataset and mitigate potential biases, we collected data from three different sources.

First, we used the dataset from Jiang et al. (2025c), in which LLMs were used to generate patent claims based on descriptions from the United States Patent and Trademark Office (USPTO). This dataset also includes human evaluations which compare the performance of different models. Second, we incorporated data from another study that investigated patent claim revision using data from the European Patent Office (EPO) and also included human evaluations (Jiang et al., 2025b). Both studies rated claims based on feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. We integrated these data to construct a comprehensive evaluation benchmark. Additionally, to further increase the dataset size and enhance robustness, we conducted new annotations by consulting patent attorneys. These additional annotations were applied to claims obtained from the aforementioned studies.

We recognize that the absolute quality scores from different sources may vary due to differences in expert interpretation. However, the relative ranking of the same claim sets should remain rather consistent across evaluations. Therefore, similar to prior work by Zuo et al. (2024), our dataset focuses on comparative evaluations. Each data point consists of a quadruplet (A, B, C, y), where A represents the reference claims, B and C are two generated claims, and the label y indicates whether B or C is better, or if they are of equal quality.

3.3 Statistics

The dataset consists of a total of 1,228 data points evaluated in five aspects. As shown in Table 3, the data distribution is relatively balanced for each aspect, with similar proportions in each category. Appendix B introduces more dataset statistics. We randomly selected 184 examples (about 15%) as the test set and used the remaining for training.

Dimension	# (B > C)	# (B = C)	# (B < C)
Completeness	426	375	427
Clarity	424	378	426
Consistency	420	386	422
Linkage	430	366	432
Quality	422	382	424

Table 3: Data distribution of the Patent-CE dataset. The data distribution is relatively balanced in each evaluation dimension.

Our benchmark offers two major advantages: **Comprehensiveness**: It incorporates patent data from multiple patent offices, which makes it more representative and robust than any previous work. **Larger scale**: Although some data builds on previous work, we manually refine and annotate more data to substantially expand the dataset size and broaden coverage.

4 Method

We propose PatClaimEval, a new automated evaluation method to assess the quality of generated patent claims compared to gold claims. Given a reference claim set P and a candidate claim set Q, the model predicts a quality score of Q, denoted as $s(Q|P)$. We train five models to evaluate patent claims from different aspects, including feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. We do not jointly train one model for all five aspects, such as using multi-task learning (Zhang and Yang, 2021), because of the conflicting optimization objectives for different tasks. For example, feature completeness and clarity are not inherently related—the claim could include all essential features, but the expression is ambiguous.

4.1 Model Architecture

We leverage Longformer[3] (Beltagy et al., 2020) as the backbone to handle long input sequences efficiently. We have not used patent-specific LLMs due to their limitations; for instance, PatentGPT is closed-source (Bai et al., 2024), and PatentGPT-J has a restricted context length (Lee, 2023). The small context length is a particular problem for

[3] https://huggingface.co/allenai/longformer-base-4096

patent texts, as it may fall short of typical patent claims. The average length of patent claims is more than 1,000 tokens (Suzgun et al., 2023), and Longformer can support up to 4,096 tokens of input. Thus, we use Longformer because it is open source, supports long input length (large enough for patent claims), and offers a controllable model size. Future work may investigate larger models with 7 or 8 billion parameters. The model encodes inputs and obtains the representation for a given input claim pair (P, Q) per

$$\mathbf{h} = \mathcal{M}([P; Q]). \tag{1}$$

It subsequently connects with a fully connected layer to get a quality score and a sigmoid function that maps the score to the range $[0, 1]$ as

$$s(Q|P) = \sigma(\mathbf{w}^\top \mathbf{h} + b). \tag{2}$$

4.2 Training Strategy

Our proposed training method draws inspiration from contrastive learning (Khosla et al., 2020) as the dataset presents relative relationships between samples. In the context of NLP, contrastive learning is used to align embeddings of related text pairs or to learn discriminative representations (Gao et al., 2021). Through contrastive loss functions, models can capture nuanced differences between text samples, making contrastive learning particularly suitable for tasks involving relative comparisons. Unlike traditional contrastive learning, which explicitly constructs positive and negative sample pairs, our method integrates label information directly to define optimization objectives tailored for different evaluation aspects.

The training data consists of quadruplets (A, B, C, y), where A is the reference claims, B and C are two generated claims, and $y \in \{1, 0, -1\}$ indicates their relative quality:

$$y = \begin{cases} 1, & \text{if } s(B|A) > s(C|A) \\ 0, & \text{if } s(B|A) = s(C|A) \\ -1, & \text{if } s(B|A) < s(C|A) \end{cases} \tag{3}$$

The model computes scores $s_B = s(B|A)$ and $s_C = s(C|A)$. To optimize the model, we define the loss function as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell(s_{B_i}, s_{C_i}, y_i), \tag{4}$$

where $\ell(s_{B_i}, s_{C_i}, y_i)$ is defined as

$$\ell = \begin{cases} ReLU(m - (s_{B_i} - s_{C_i})), & \text{if } y_i = 1, \\ ReLU(|s_{B_i} - s_{C_i}| - n), & \text{if } y_i = 0, \\ ReLU\left(m - (s_{C_i} - s_{B_i})\right), & \text{if } y_i = -1, \end{cases} \tag{5}$$

where m is the margin hyper-parameter that enforces a minimum separation between scores for distinct quality levels, and n is a tolerance parameter that allows small differences between scores when the two claim sets are judged equally good.

By minimizing this loss, the model learns to align the predicted scores with the relative quality judgments. The margin m is a hyper-parameter that controls the separation between scores, ensuring that the model is confident in its predictions for cases where one claim set is clearly better than the other. This objective function allows the model to capture fine-grained distinctions in quality across diverse claim pairs. We introduce the training and evaluation details in Appendix C.

5 Experiments

5.1 Baselines

BLEU (Papineni et al., 2002) and **ROUGE** (Lin, 2004) are classic metrics widely used for evaluating text overlaps. BLEU measures n-gram precision by comparing candidate and reference texts, while ROUGE primarily evaluates recall-based overlap, commonly used in summarization. **METEOR** (Banerjee and Lavie, 2005) improves on BLEU by incorporating synonymy, stemming, and other linguistic features, thereby providing a more flexible approach to measuring textual overlap.

BERTScore (Zhang et al., 2019) computes similarity using contextualized embeddings from BERT (Devlin et al., 2019), enabling a more nuanced assessment of semantic similarity between reference and candidate sentences. **BARTScore** (Yuan et al., 2021), derived from the BART model (Lewis et al., 2020), uses a generative scoring approach to evaluate the likelihood of generating the candidate text from a reference. **MoverScore** (Zhao et al., 2019) measures the semantic similarity by calculating the minimum cost of transforming candidate embeddings to reference embeddings, effectively capturing semantic alignment. **SimCSE** (Gao et al., 2021) further enhances representation quality by using contrastive learning to generate sentence embeddings, which have been shown to perform well in semantic similarity tasks.

We also test recent multi-dimensional evaluation frameworks, including **UniEval** (Zhong et al., 2022) and **AlignScore** (Zha et al., 2023). UniEval provides a unified evaluation protocol for different aspects of natural language generation, such as coherence and fluency. Since our dataset does not include context information or source texts, we use UniEval to evaluate the relevance between generated and reference claims. Additionally, we use AlignScore as a representative to assess factual consistency between source and generated content.

The LLM-as-a-judge paradigm is becoming popular (Zheng et al., 2023), where LLMs are used as evaluators of generated content. This approach leverages the capabilities of pre-trained LLMs, such as GPT-4 (OpenAI, 2023), to serve as surrogate judges that can assess generated text for qualities like fluency, coherence, and factual consistency. In our experiments, we specifically focus on **G-Eval-4** (Liu et al., 2023) because it has shown high agreement with human preference across multiple benchmarks (Zheng et al., 2023). Other LLM-as-a-judge models are not tested because they use synthetic examples generated by GPT-4 for training, such as JudgeLM (Zhu et al., 2023) PandaLM (Wang et al., 2024). We ask GPT-4 to evaluate the given claims through chain-of-thought (CoT) prompting (Wei et al., 2022) by comparing them to the reference claims. The evaluation dimensions are the same as human expert metrics. We introduce detailed settings in Appendix D.

5.2 Evaluations

We used the **Kendall-Tau correlation** to assess the overall alignment with human judgment, following the approach of previous work by Zuo et al. (2024). This correlation metric evaluates the consistency of the global ranking while disregarding minor errors in individual predictions. We additionally report the **Spearman correlation**. Compared to Kendall-Tau, Spearman is more sensitive to large rank differences, providing a complementary perspective on the metric ability to predict relative claim quality.

Since the dataset originally presents a three-way classification problem, we also use **accuracy** and **F1 scores** to assess model performance. These metrics reflect the model's ability to make precise decisions for individual input pairs, providing a more comprehensive view of its effectiveness. Classification labels can be obtained directly from G-Eval-4, while for other metrics, we assume quality scores to be equivalent if the score differences are less than 10^{-4}.

6 Results

6.1 Correlations with Human Evaluations

Table 4 presents the Kendall-Tau and Spearman correlation between different automated metrics and human evaluation results across five criteria: feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality.

Overall, **PatClaimEval demonstrates the highest correlation with human evaluations across all criteria**, suggesting its effectiveness in evaluating patent claim quality. For feature completeness, PatClaimEval achieves a correlation of $\tau = 0.400$ and $\rho = 0.504$, which outperforms all other metrics. This finding holds consistently across other criteria, with correlations of $\tau = 0.461$ and $\rho = 0.518$ for clarity, $\tau = 0.354$ and $\rho = 0.424$ for consistency, $\tau = 0.419$ and $\rho = 0.518$ for linkage, and $\tau = 0.477$ and $\rho = 0.602$ for overall quality. Notably, these values are not only the highest but also significantly surpass existing metrics in their alignment with human judgments. Particularly in overall quality, PatClaimEval outperforms the second-best result by approximately 41.5% and 58.0% for Kendall-Tau and Spearman correlation respectively.

In addition, **N-gram-based metrics demonstrate relatively higher correlations than embedding-based methods in evaluating patent claims.** While N-gram-based methods can sometimes achieve correlation scores exceeding 0.3 in different evaluation aspects, embedding-based metrics rarely surpass this threshold. For instance, ROUGE-L achieves the second-highest Spearman correlation in logical linkage with a score of 0.391. N-gram-based methods rely on surface-level overlap between generated and reference text, without capturing semantic information or contextual relevance. These methods typically underperform compared to embedding-based approaches, which calculate semantic similarities, in standard text evaluation tasks (Zhang et al., 2019; Zhao et al., 2019; Yuan et al., 2021). However, patent claim evaluation results diverge from these prior findings due to its unique focus on patent examination criteria. Both reference and candidate claims describe the same invention but often use different expressions. In this context, high semantic similarity does not necessarily indicate adherence to patent require-

Type	Metric	Completeness		Clarity		Consistency		Linkage		Quality	
		τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
N-gram	BLEU-1	0.305	0.345	0.359	0.401	0.284	0.329	0.335	0.376	0.326	0.369
	BLEU-4	0.271	0.304	0.280	0.312	0.227	0.263	0.256	0.289	0.269	0.305
	ROUGE-1	0.305	0.342	0.314	0.351	0.238	0.279	0.301	0.341	0.292	0.332
	ROUGE-2	0.305	0.342	0.280	0.312	0.215	0.251	0.268	0.303	0.269	0.306
	ROUGE-L	0.282	0.317	0.280	0.312	0.261	0.303	0.346	0.391	0.303	0.344
	METEOR	0.316	0.358	0.371	0.414	0.307	0.355	0.324	0.364	0.292	0.331
Embedding	BERTScore	0.241	0.279	0.251	0.281	0.242	0.283	0.272	0.303	0.239	0.268
	BARTScore	0.165	0.188	0.130	0.146	0.211	0.242	0.196	0.219	0.164	0.185
	MoverScore	0.199	0.227	0.199	0.217	0.223	0.264	0.231	0.265	0.210	0.243
	SimCSE	0.177	0.196	0.165	0.173	0.143	0.165	0.220	0.246	0.165	0.186
Miscellaneous	UniEval	0.339	0.383	0.337	0.375	0.261	0.302	0.301	0.338	0.337	0.381
	AlignScore	0.146	0.162	0.145	0.160	0.261	0.305	0.200	0.226	0.224	0.255
LLM-as-a-judge	G-Eval-4	0.377	0.410	0.412	0.481	0.276	0.353	0.350	0.385	0.277	0.310
Ours	PatClaimEval	**0.400**	**0.504**	**0.461**	**0.518**	**0.354**	**0.424**	**0.419**	**0.518**	**0.477**	**0.602**

Table 4: Kendall-Tau (τ) and Spearman (ρ) correlation of automated metrics with human evaluation results. The highest number in each criterion is in **bold**, and the second-best result is underlined. PatClaimEval shows the highest correlation with human assessments in all criteria.

ments, resulting in weak correlations with human judgments. In contrast, gold-standard patent claims use precise language and expressions designed to meet examination standards. Thus, more overlaps with these gold claims may better reflect higher quality, potentially explaining why simple overlap-based methods outperform embedding-based similarity approaches in this domain. These findings extend to metrics of UniEval and AlignScore. While AlignScore assesses factual consistency and shows less correlation, UniEval that measures relevance between candidate and reference claims performs relatively better.

G-Eval-4 shows strong performance in evaluating completeness, clarity, and linkage. G-Eval-4 achieves correlation scores of $\tau = 0.377$ and $\rho = 0.410$ for completeness, which surpasses all other metrics except for PatClaimEval and is consistent with findings from prior research (Jiang et al., 2025b). The high correlation in feature completeness can be attributed to GPT-4's proven capabilities in information extraction (OpenAI, 2023; Li et al., 2023). In the context of claim evaluation, GPT-4 effectively extracts key features from both reference and candidate claims. In consequence, it enables accurate comparisons and reaches high scores in feature completeness assessments. However, its performance in terminology consistency and overall quality is less impressive. A plausible explanation is that GPT-4 is not extensively trained on patent-specific texts, which limits its ability to comprehend the unique linguistic and structural requirements of patent claims. Consequently, relying solely on prompting without further fine-tuning may be insufficient for accurately evaluating patent claims.

6.2 Classification Performance

Table 5 presents the accuracy and F1 scores of different metrics on each evaluation criterion as a classification problem, including feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality.

PatClaimEval achieves the highest accuracy and F1 scores across nearly all evaluation criteria. Specifically, for conceptual clarity, PatClaimEval achieves an accuracy of 60.3% and an F1 score of 59.5%, outperforming all other metrics. This superior performance extends to consistency and overall quality, where PatClaimEval consistently outperforms other methods. In feature completeness, G-Eval-4 demonstrates slightly better performance to PatClaimEval, with both accuracy and F1 scores of 54.8%. Despite these strengths, PatClaimEval's absolute scores in some evaluation criteria, such as consistency, remain modest (50.0% accuracy and 49.3% F1 score). The moderate absolute scores indicate potential for improvement, such as expanding dataset sizes, larger models, or more sophisticated training strategies. Nonetheless, PatClaimEval represents a significant advancement as it achieves a 3.8% improvement in accuracy and a 10.5% increase in F1 score over the second-best method for overall quality evaluation. It currently

Type	Metric	Completeness		Clarity		Consistency		Linkage		Quality	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
N-gram	BLEU-1	50.5	42.6	54.3	46.7	47.8	39.2	53.8	46.5	52.2	44.3
	BLEU-4	48.9	41.4	50.5	43.5	45.1	37.0	50.0	43.2	49.5	42.0
	ROUGE-1	50.5	42.8	52.2	44.8	45.7	37.5	52.2	45.0	50.5	42.9
	ROUGE-2	50.5	42.8	50.5	43.4	44.6	36.6	50.5	43.6	49.5	42.0
	ROUGE-L	49.5	41.8	50.5	43.5	46.7	38.3	54.3	46.9	51.1	43.4
	METEOR	51.1	43.2	54.9	47.2	48.9	40.1	53.3	46.0	50.5	42.8
Embedding	BERTScore	46.7	39.1	48.9	42.2	45.7	37.6	51.1	44.8	48.4	41.8
	BARTScore	43.5	35.9	42.9	36.3	44.6	36.4	47.3	40.7	44.6	37.5
	MoverScore	45.1	38.4	46.7	41.1	44.6	36.8	48.4	42.0	46.2	39.5
	SimCSE	44.6	38.4	45.7	40.7	41.3	34.6	48.4	42.6	44.6	38.6
Miscellaneous	UniEval	52.2	44.1	53.3	45.8	46.7	38.3	52.2	45.0	52.7	44.8
	AlignScore	42.9	36.4	44.0	38.0	46.7	38.3	47.3	40.8	47.3	40.1
LLM-as-a-judge	G-Eval-4	**54.8**	**54.8**	55.6	55.9	45.9	43.7	**54.8**	54.6	49.6	46.9
Ours	PatClaimEval	52.7	53.2	**60.3**	**59.5**	**50.0**	**49.3**	52.7	**54.7**	**56.5**	**57.4**

Table 5: Accuracy (Acc %) and F1 scores (F1 %) on each evaluation criterion. The highest number in each column is in **bold**, and the second-best result is underlined. PatClaimEval demonstrates relatively high and balanced accuracy and F1 scores across all evaluation criteria.

stands as the most effective approach for patent claim evaluation.

PatClaimEval and G-Eval-4 exhibit balanced performance between accuracy and F1 scores. Both models achieve similar accuracy and F1 scores across all five evaluation criteria, in contrast to other metrics, where F1 scores are normally notably lower than their accuracies. This balance reflects an effective trade-off between precision and recall. Although G-Eval-4 does not lead in accuracy across all aspects, its F1 scores are consistently higher than other metrics except for PatClaimEval. Based on a careful examination of the results, we attribute this strength to G-Eval-4's ability to handle "equal cases", in which two candidate claims receive identical quality scores. Metrics such as N-gram-based and embedding-based methods struggle to evaluate such cases effectively, resulting in discrepancies between their accuracy and F1 scores. The balanced performance of PatClaimEval and G-Eval-4 highlights their robustness and reliability in patent claim evaluation.

6.3 Qualitative Analysis

We show an example of claim comparison in Table 7 to demonstrate the inherent challenges of this task, where A represents the gold claim, and B and C are candidate claims. We identify three types of differences between generated claims B and C. First, Claim C demonstrates higher clarity and language precision. It correctly uses *an annular edge*, whereas Claim B incorrectly uses *a annular edge*, a basic grammatical error. Furthermore, in Claim 3, C uses *further comprising*, which aligns with the gold claim and drafting conventions, while B uses the inappropriate *comprises*. Second, Claim C exhibits a stronger logical linkage between components. It introduces dependent clauses in Claim 3 properly using *wherein* that preserves structural relationships between features, whereas Claim B omits such linkages. Third, Claim C uses the phrase *are configured to* when describing some features. While this phrasing deviates from the gold claim, it does not degrade the quality. Overall, Claim C is better than Claim B. However, current metrics cannot capture such subtle and special differences, which could lead to unreliable performance in claim evaluation.

7 Conclusion

We introduce Patent-CE, the first comprehensive benchmark for evaluating patent claims. Patent-CE includes comparative evaluations annotated by patent experts, which focus on five key criteria that align with established patent examination standards: feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. Moreover, we propose PatClaimEval, a novel multi-dimensional evaluation method specifically designed for patent claims. Extensive experiments demonstrate the effectiveness of PatClaimEval. It achieves the highest correlation with human expert evaluations across all assessment criteria when compared to existing metrics.

This research provides valuable resources for developing automated evaluation methods of patent claims and establishes a solid foundation for more reliable assessments of claim generation systems.

Limitations

We acknowledge several limitations in this research. Firstly, the dataset used in this study includes only patents documented in English, which may affect the applicability to patents in other languages. In addition, the correlations of our method with human assessments are still somewhat low (Kendall-Tau < 0.5) and improvements are needed. Better semantic methods or different kinds of CoT prompting strategies may also be worth investigating. Furthermore, our evaluation approach relies on a gold standard. It provides a more reliable way to evaluate patent claims and is especially useful when developing related models for claim generation. However, real-world patent examinations by patent offices consider a range of criteria, including novelty, non-obviousness, and language requirements, without necessarily referencing a predefined gold standard. Patents are evaluated based on their intrinsic merit and their relation to prior art. Therefore, exploring reference-free evaluation approaches for patent claims is an important and worthwhile direction for future work.

Ethics Statement

GPT-4 is under a commercial license provided by OpenAI, and we access it through its API. The use of existing artifacts, including models, evaluation metrics, and datasets, is consistent with their intended use. Our proposed dataset is used for patent claim generation evaluation and will be released under *CC-BY-NC-4.0* license. This dataset does not include potential personal information or offensive content, and no ethics review board was involved.

References

Zilong Bai, Ruiji Zhang, Linqing Chen, Qijun Cai, Yuan Zhong, Cong Wang, Yan Fang, Jie Fang, Jing Sun, Weikuan Wang, et al. 2024. Patentgpt: A large language model for intellectual property. *arXiv preprint arXiv:2404.18255*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Patent Office. 2000. Epc - the european patent convention. `https://www.epo.org/en/legal/epc/2020/regulations.html`. Accessed: 2023-06-12.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Robert C Faber. 1990. *Landis on mechanics of patent claim drafting*. Practising Law Institute New York.

M. Frumkin. 1947. Early history of patents for innovation. *Transactions of the Newcomen Society*, 26(1):47–56.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lekang Jiang and Stephan M Goetz. 2025. Natural language processing in the patent domain: a survey. *Artificial Intelligence Review*, 58(7):214.

Lekang Jiang, Chengzu Li, and Stephan Goetz. 2025a. Enriching patent claim generation with european patent dataset. *arXiv preprint arXiv:2505.12568*.

Lekang Jiang, Pascal A. Scherz, and Stefan Goetz. 2025b. Patent-CR: A dataset for patent claim revision. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2300–2314, Albuquerque, New Mexico. Association for Computational Linguistics.

Lekang Jiang, Caiqi Zhang, Pascal A. Scherz, and Stefan Goetz. 2025c. Can large language models generate high-quality patent claims? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1272–1287, Albuquerque, New Mexico. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Jieh-Sheng Lee. 2020. Controlling patent text generation by structural metadata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3241–3244.

Jieh-Sheng Lee. 2023. Evaluating generative patent language models. *World Patent Information*, 72:102173.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Cislo & Thomas LLP. 2023. Typical fees. Accessed: 2024-10-15.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Adam Mossoff. 2000. Rethinking the development of patents: an intellectual history, 1550-1800. *Hastings Lj*, 52:1255.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott D Kominers, and Stuart Shieber. 2023. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *Advances in neural information processing systems*, 36:57908–57946.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213,

Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

You Zuo, Kim Gerdes, Éric Clergerie, and Benoît Sagot. 2024. PatentEval: Understanding errors in patent generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2687–2710, Mexico City, Mexico. Association for Computational Linguistics.

A Human Annotations

We invite licensed patent attorneys for human evaluations. These professionals are provided with reference claims and candidate claims for assessment. They are informed about the intended use of the evaluation results. Table 8 outlines the detailed evaluation criteria, aligned with prior research (Jiang et al., 2025c). We compare the scores and construct the comparative evaluation dataset.

B Dataset Statistics

We report the token length statistics of the Patent-CE dataset using the Longformer tokenizer in this section. The results are summarized as follows: the minimum length is 156 tokens, the maximum length is 1,461 tokens, the average length is 644 tokens, the median length is 631 tokens, and the standard deviation is 245 tokens. All claims fall within the token limit of Longformer (4096 tokens), and thus no truncation or segmentation strategies were used. This ensures that input length limitations do not affect the evaluation results. Since the dataset does not include very long claims, the proposed method may not generalize well to extremely long claims that exceed the model's input capacity.

C Experimental Details

All training and testing processes are conducted on NVIDIA A100 GPUs. The total running time is about 20 hours. We randomly select 10% samples from the training set as the validation set. During training, we use a batch size of 4, a learning rate of 5e-6, a weight decay of 0.01, and training epochs of 10. For BLEU, ROUGE, METEOR, and BERTScore, we use the package from the HuggingFace *evaluate* library.[4] For MoverScore[5], BARTScore[6], AlignScore[7], SimCSE[8], and UniEval[9], we use their code from original repositories. We use the *scipy* Python library to calculate the correlation scores and *scikit-learn* for accuracy and F1 scores.

D G-Eval-4

We use the following prompt for G-Eval consistent with previous research (Jiang et al., 2025b), as shown in Table 6. We use GPT-4 to evaluate feature completeness, conceptual clarity, terminology consistency, and logical linkage. The overall quality is calculated based on the same formula of human evaluation in Table 8.

[4] https://github.com/huggingface/evaluate
[5] https://github.com/AIPHES/emnlp19-moverscore
[6] https://github.com/neulab/BARTScore
[7] https://github.com/yuh-zha/AlignScore
[8] https://github.com/princeton-nlp/SimCSE
[9] https://github.com/maszhongming/UniEval

Instructions:
You will be given the draft claims and the referenced claims of the same patent. Your task is to rate the draft claims on four metrics using the referenced claims as the gold standard. Please make sure you read and understand these instructions carefully. Keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
1. Completeness of Essential Features (0–100)
The extent to which the generated claims encapsulate all critical aspects of the invention.
0–20: Most essential features are missing or poorly described.
21–40: Some essential features are present but significant gaps remain.
41–60: Majority of essential features are covered but with minor omissions.
61–80: Almost all essential features are well described with very few gaps.
81–100: All essential features are thoroughly and comprehensively covered.

2. Conceptual Clarity (0–100)
The clarity and unambiguity of the language used in the claims.
0–20: Claims are very unclear and ambiguous.
21–40: Claims have significant clarity issues, making them difficult to understand.
41–60: Claims are mostly clear but contain some ambiguous language.
61–80: Claims are clear with minimal ambiguity.
81–100: Claims are exceptionally clear and completely unambiguous.

3. Consistency in Terminology (0–100)
The uniformity in the use of terms throughout the claims.
0–20: Terminology is highly inconsistent.
21–40: Significant inconsistencies in terminology.
41–60: Some inconsistencies in terminology but mostly uniform.
61–80: Terminology is largely consistent with minor inconsistencies.
81–100: Terminology is completely consistent throughout.

4. Technical Correctness of Feature Linkages (0–100)
The accuracy with which the features are interconnected and related.
0–20: Features are poorly linked with many inaccuracies.
21–40: Significant issues with the linkages of features.
41–60: Mostly accurate linkages with some incorrect connections.
61–80: Accurate linkages with minor inaccuracies.
81–100: Features are accurately and correctly linked throughout.

Evaluation Steps:
1. Read the referenced claims carefully and identify the invention's features. Assume the referenced claims have scores of 100 in all Evaluation Criteria.
2. Read the draft claims and compare them to the referenced claims.
3. Assign a score for each metric based on the Evaluation Criteria.

Example:
Referenced Claims: «`Claims`»
Draft Claims: «`Claims`»
Evaluation Form (scores ONLY):
- Completeness of Essential Features: X
- Conceptual Clarity: X
- Consistency in Terminology: X
- Technical Correctness of Feature Linkages: X

Table 6: G-Eval prompt used for claim evaluation originated from Jiang et al. (2025b)

E Example Claim Comparison

We present an example of claim comparison in Table 7, where the differences between Claim B and C are marked in blue.

Gold Claim A

1. A shroud for connecting to a container having a closure portion, the shroud comprising: a housing having a luer connector; a spike having a fluid lumen transitioning into the connector; a plurality of segments terminating in a continuous annular edge surrounding the spike and defining a plurality of openings; and a plurality of protrusions circumferentially spaced, each of the plurality of protrusions having (i) a proximal end connected to the continuous annular edge and (ii) a distal end positioned in one of the plurality of openings, the distal end of each protrusion inwardly extending from the continuous annular edge, the protrusions defining a receiving volume dimensioned to receive the closure portion of the container, whereby the protrusions deflect outwardly away from an initial position during contact and traversing of the closure portion of the container and return to approximately their initial position when the closure portion of the container passes the distal end of the protrusion, whereby the protrusions deflect inwardly towards the closure portion of the container when a removing force is applied to the container, thereby preventing removal of the container.
2. The shroud of claim 1, wherein the housing comprises a surface with parallel raised features for facilitating gripping by a user.
3. The shroud of claim 1, further comprising a vent housing and wherein the spike further comprises a vent lumen connected to the vent housing.
4. The shroud of claim 3, wherein the vent housing further comprises a filter.

Generated Claim B

A shroud for connecting to a container having a closure portion, the shroud comprising: a housing having a connector; a spike having a fluid lumen fluidically coupled to the connector; a plurality of segments terminating in a annular edge surrounding the spike and defining a plurality of openings; a plurality of protrusions circumferentially spaced and inwardly extending from the annular edge, the protrusions defining a receiving volume dimensioned to receive the closure portion of the container, wherein the protrusions have a proximal end connected to the annular edge and a distal end positioned in the opening, whereby the protrusions deflect outwardly away from an initial position during contact and traversing of the closure portion of the container and return to approximately their initial position when the closure portion of the container passes the distal end of the protrusion, whereby the protrusions deflect inwardly towards the closure portion of the container when a removing force is applied to the container, thereby preventing removal of the container.
2. The shroud of claim 1, wherein the housing comprises a surface with raised features for facilitating gripping by a user.
3. The shroud of claim 1 further comprises a vent housing and the spike further comprises a vent lumen coupled to the vent housing.
4. The shroud of claim 3, wherein the vent housing further comprises a filter.

Generated Claim C

1. A shroud for connecting to a container having a closure portion, the shroud comprising: a housing having a connector; a spike having a fluid lumen fluidically coupled to the connector; a plurality of segments terminating in an annular edge surrounding the spike and defining a plurality of openings; a plurality of protrusions circumferentially spaced and inwardly extending from the annular edge, the protrusions defining a receiving volume dimensioned to receive the closure portion of the container, wherein the protrusions have a proximal end connected to the annular edge and a distal end positioned in the opening, whereby the protrusions are configured to deflect outwardly away from an initial position during contact and traversing of the closure portion of the container and return to approximately their initial position when the closure portion of the container passes the distal end of the protrusion, whereby the protrusions are further configured to deflect inwardly towards the closure portion of the container when a removing force is applied to the container, thereby preventing removal of the container.
2. The shroud of claim 1, wherein the housing comprises a surface with raised features for facilitating gripping by a user.
3. The shroud of claim 1, further comprising a vent housing and wherein the spike further comprises a vent lumen coupled to the vent housing.
4. The shroud of claim 3, wherein the vent housing further comprises a filter.

Table 7: An example of claim comparison. Differences between B and C are marked in blue, and C is better.

Criteria	Rating Description
Feature Completeness	• **0-2:** Most essential features are missing or poorly described. • **3-4:** Some essential features are present but significant gaps remain. • **5-6:** Majority of essential features are covered but with minor omissions. • **7-8:** Almost all essential features are well described with very few gaps. • **9-10:** All essential features are thoroughly and comprehensively covered.
Conceptual Clarity	• **0-2:** Claims are very unclear and ambiguous. • **3-4:** Claims have significant clarity issues, making them difficult to understand. • **5-6:** Claims are mostly clear but contain some ambiguous language. • **7-8:** Claims are clear with minimal ambiguity. • **9-10:** Claims are exceptionally clear and completely unambiguous.
Terminology Consistency	• **0-2:** Terminology is highly inconsistent. • **3-4:** Significant inconsistencies in terminology. • **5-6:** Some inconsistencies in terminology but mostly uniform. • **7-8:** Terminology is largely consistent with minor inconsistencies. • **9-10:** Terminology is completely consistent throughout.
Logical Linkages	• **0-2:** Features are poorly linked with many inaccuracies. • **3-4:** Significant issues with the linkages of features. • **5-6:** Mostly accurate linkages with some incorrect connections. • **7-8:** Accurate linkages with minor inaccuracies. • **9-10:** Features are accurately and correctly linked throughout.
Overall Quality	• Calculated by: $(completeness * 4 + clarity * 2 + consistency * 2 + correctness * 3) \div 11$ • **0-2:** Very poor overall quality, fails to meet most criteria. • **3-4:** Low overall quality with significant issues across multiple criteria. • **5-6:** Average overall quality, meets criteria at a basic level. • **7-8:** High overall quality with minor issues. • **9-10:** Excellent overall quality, meets or exceeds all criteria.

Table 8: Rating criteria for human annotation deriving from Jiang et al. (2025c)