Thinking Silently for Patent Claim Generation

Lekang Jiang[1], **Wenjun Sun**[123], **Stephan Goetz**[1]
[1]University of Cambridge [2]National Science Library, Chinese Academy of Sciences
[3]Department of Information Resources Management, School of Economics and Management,
University of Chinese Academy of Sciences
{lj408, ws462, smg84}@cam.ac.uk

Abstract

This work investigates whether Chain-of-Thought (CoT) prompting benefits patent claim generation. We demonstrate a counter-intuitive but important empirical finding: implicit CoT, where reasoning is kept internal rather than explicitly verbalized, consistently outperforms explicit CoT. Through systematic analysis, we show that explicit CoT can introduce an unnecessary information bottleneck for claim generation. Verbalized reasoning may compromise the quality of final outputs through three specific mechanisms: abstraction of critical details, disruption of internalized generation patterns, and cascading error propagation. Our findings provide new insights into structured text generation and CoT applications.

1 Introduction

Automated patent claim drafting represents a significant challenge in natural language processing (NLP) (Zhao et al., 2023; Jiang and Goetz, 2025), because it demands not only exceptional linguistic proficiency but also stringent requirements on logical reasoning, strict adherence to structural conventions, and the integration of domain-specific knowledge (Jiang et al., 2025d; Wang et al., 2024).

To unlock and guide the deeper reasoning capabilities of large language models (LLMs), the Chain-of-Thought (CoT) prompting technique has emerged as a powerful method (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Zhang et al., 2023). CoT encourages models to break down complex problems into a series of intermediate reasoning steps before arriving at a final answer. In consequence, it could significantly enhance performance on arithmetic, commonsense, and symbolic reasoning tasks.

However, recent studies have reported that the effectiveness of CoT in some tasks, such as pattern-based and patent-related tasks, is limited or even counterproductive (Zheng et al., 2025; Wang et al., 2025). These findings raise critical questions about whether CoT can benefit patent claim drafting.

In this work, we propose a CoT method for claim generation. Importantly, our results reveal a counter-intuitive finding: implicit CoT consistently outperforms explicit CoT across multiple models, with explicit verbalization degrading claim quality on all evaluation dimensions. Through diagnostic experiments and case studies, we trace the decrease to an information bottleneck created by reasoning externalization, which manifests through three mechanisms: errors cascade irrecoverably across reasoning steps, verbalized reasoning abstracts away legally critical details, and the explicit reasoning phase disrupts the model's internalized generation patterns.

2 Related Work

2.1 CoT Prompting

Wei et al. (2022) introduced CoT prompting, which provides LLMs with few-shot exemplars that include intermediate reasoning steps to improve the models' performance on complex arithmetic, commonsense, and symbolic reasoning tasks. This finding suggested that eliciting a step-by-step reasoning process is key to unlocking the latent capabilities of LLMs. Subsequent work by Kojima et al. (2022) revealed that this reasoning ability could be unlocked even without few-shot exemplars. Their zero-shot CoT approach merely appends a simple, generic instruction, such as "Let's think step by step", to the prompt. This discovery was pivotal, as it showed that the reasoning mechanism is an emergent capability of sufficiently scaled models, which can be triggered by a direct instruction rather than imitation. More complex reasoning structures have also been explored, such as the Tree of Thoughts (ToT) (Yao et al., 2023) and the Graph of Thoughts (GoT) (Besta et al., 2024). However, recent studies have reported that the effectiveness of CoT in some

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 329–341
November 6-10, 2026 ©2026 Association for Computational Linguistics

tasks, including patent-related tasks, is limited or even counterproductive (Zheng et al., 2025; Wang et al., 2025).

2.2 Patent Claim Generation

Drafting high-quality patent claims is a time-consuming, costly, and skill-intensive process (European Patent Office, 2000). Thus, recent research has explored LLMs to assist inventors and patent professionals in automatically generating well-structured claims (Jiang and Goetz, 2025). The research focus has gradually shifted from earlier approaches that train neural models on patent abstracts (Lee and Hsiang, 2020; Lee, 2020) towards using LLMs to generate claims from detailed patent descriptions (Jiang et al., 2025d; Wang et al., 2024). As descriptions provide richer technical context, these methods can significantly improve the completeness and precision of the generated claims.

3 Experiments

3.1 Task

The task aims to automatically generate a set of high-quality patent claims C based on a given technical description D of an invention: $C = G(I, D)$, where G represents LLMs, and I is the instruction prompt. Unlike conventional text generation tasks such as summarization or translation, generating patent claims presents a significant challenge that demands deep semantic understanding and complex logical reasoning. The model must distill the core inventive concepts from a lengthy technical document and synthesize these concepts into a hierarchically structured set of claims. Each dependent claim must logically narrow the scope of its parent claim and adhere to strict legal formalism. Furthermore, every element within the claims must be fully supported by and grounded in the provided description. Therefore, an intermediate reasoning process is recommended before writing final claims to identify essential invention features, structure features logically, and articulate claims formally.

3.2 CoT Method

Although CoT has significantly advanced the general reasoning capabilities of LLMs, it often relies on either manually crafted exemplars (Wei et al., 2022) or generic instructions (Kojima et al., 2022) that may not be optimal for highly specialized, domain-specific tasks that demand a precise logical

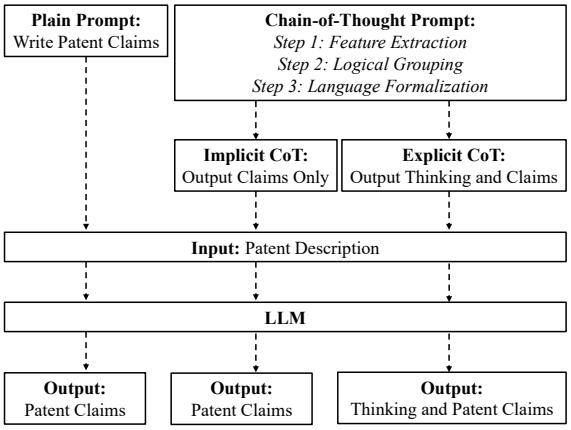


Figure 1: Overview of the plain and CoT prompting framework for patent claim generation.

flow. Our work builds upon the principles of zero-shot CoT (Kojima et al., 2022) and addresses the domain gap. In contrast to the use of a generic trigger phrase, we introduce a structured zero-shot CoT approach: $C = G(I_{CoT}, D)$ as shown in Figure 1. Our approach decomposes the generation process into three structured steps to simulate the professional drafting process. Completed CoT prompts are illustrated in Appendix Table 4.

Step 1: Feature Extraction. The model first identifies the key technical features in patent descriptions and their importance ranking.

Step 2: Logical Grouping. The extracted features are organized logically based on their functional relationships and dependencies.

Step 3: Language Formalization. Organized features are written into textual claims using formal patent language to ensure clarity, consistency, and technical accuracy.

By providing the model with a task-specific framework via structured instructions, we guide its reasoning process to align with the domain-specific requirements of patent claim drafting. This method maintains the simplicity of the zero-shot approach but imposes a necessary structure to enhance the completeness, logical coherence, and language formality of the generated output.

We experiment with the plain prompt and different CoT prompting variants (Figure 1):

Plain Prompt: the model outputs only the final claims without three-step CoT prompting: $C = G(I, D)$.

Implicit CoT: the model is guided to think step-by-step internally and only outputs the final claims: $C = G(I_{CoT_Implicit}, D)$.

Explicit CoT: the intermediate reasoning steps

are explicitly written out prior to claim generation: $R, C = G(I_{CoT_Explicit}, D)$, where R is the reasoning process.

3.3 Dataset

Our main experiments are conducted on the challenging subset of the EPD dataset (Jiang et al., 2025a), which contains 693 English-language patents from the European Patent Office (EPO). We select this dataset because EPD features high-quality granted patents, and the challenging subset better reflects real-world drafting scenarios, which enables a more rigorous evaluation of LLM performance. To examine generalization, we also test on HUPD-DCG (Jiang et al., 2025d), which includes 1,311 US patents. Results are consistent with those on EPD and are presented in Appendix E. Additional dataset details are provided in Appendix B.

3.4 Models

We mainly focus on Llama-3.1-8B (Dubey et al., 2024), Mixtral-8×7B (Jiang et al., 2024), and GPT-4o (OpenAI, 2024) in the paper to explore the effect of model sizes and architectures. We further report the results of Mistral-7B (Jiang et al., 2023) and Llama-3.1-70B (Dubey et al., 2024) in Appendix E.

3.5 Evaluation Metrics

We mainly use human evaluation for our experiments for high precision, and report automatic evaluation in Appendix. We invite licensed patent attorneys, who follow standard patent-examination guidelines and assess generated claims against gold claims across three dimensions (0–100 scale): (1) Content Coverage, (2) Logical Structure, and (3) Language Quality. Due to the difficulty and time intensity of expert assessment, we evaluate 120 randomly selected examples. Details and evaluation guidelines are provided in Appendix C.

4 Results and Analysis

Table 1 presents human expert evaluation results across different models. We also provide a representative case study in Table 2 and include full model outputs in Appendix Table 10.

4.1 Main Results

Our results show that implicit CoT prompting can improve claim generation quality. While we agree that the gains for smaller models are modest, the improvement for larger models such as GPT-4o is both substantial and stable, with higher evaluation scores and smaller standard deviations. As shown in Table 1, GPT-4o with implicit CoT achieves an average increase of 5.2 points in human evaluations, with the most notable gain observed in logical structure (5.9 points). The improvements can be attributed to two key factors. First, CoT decomposes claim generation into three sub-tasks that clarify task objectives at each stage, reduce cognitive load, and guide models to generate better results. Second, this decomposition activates the model's latent reasoning, planning, and domain knowledge capabilities, which may remain underutilized in direct generation.

Model / Setting	Content	Logic	Language	Avg
Llama-3.1-8B	82.3±2.4	82.5±2.3	81.2±2.4	82.0
w/ Implicit CoT	82.7±2.2	82.5±2.1	81.8±2.1	82.3
w/ Explicit CoT	82.4±2.2	82.5±2.3	81.4±2.0	82.1
Mixtral-8×7B	84.6±2.3	85.0±2.4	84.6±2.1	84.7
w/ Implicit CoT	85.4±2.0	85.4±2.1	85.8±2.0	85.5
w/ Explicit CoT	84.6±2.2	85.2±2.2	84.7±2.1	84.8
GPT-4o	88.5±2.4	86.3±2.5	87.7±1.8	87.5
w/ Implicit CoT	**93.0±1.5**	**92.2±2.1**	**92.8±1.6**	**92.7**
w/ Explicit CoT	90.0±1.8	89.8±2.3	91.7±1.6	90.5

Table 1: Human evaluation results on different models.

However, we observe that explicit CoT substantially degrades performance compared to implicit CoT across all human evaluation criteria. As shown in Table 1, explicit CoT for GPT-4o leads to consistent drops across all human evaluation dimensions, with an average decrease of 2.2 points. This result contrasts with prior findings in arithmetic or logic reasoning tasks, where explicit CoT typically improves model accuracy (Wei et al., 2022; Kojima et al., 2022; Yu, 2024).

4.2 The Bottleneck of Explicit CoT

The consistent underperformance of explicit CoT demands explanation. We argue that the root cause lies in a fundamental mismatch between what explicit CoT does and what patent claim generation requires. In reasoning tasks such as arithmetic, intermediate steps can be expressed in natural language with minimal information loss—each step is a self-contained, verifiable proposition. Patent claim generation is qualitatively different: the "reasoning" involved—selecting which technical details are legally critical, weighing alternative phrasings, planning hierarchical claim structure—is deeply entangled and hard to be faithfully linearized into text. When the model is forced to externalize this process, the resulting reasoning text acts as an *information bottleneck*: the model's rich

Features	Gold	Standard Prompt	Implicit CoT	Explicit CoT	
	Claims	Claims	Claims	Thinking	Claims
Core Components					
Air-tight container, Injector, Intraocular Lens (IOL), Water reservoir	✓	✓	✓	✓	✓
Container Details					
Comprises a tray, Side walls, Bottom walls	✓	✗	✓	✗	✓
IOL Details					
Foldable	✓	✓	✓	✓	✓
Hydrophilic	✓	✗	✗	✗	✗
Unfolded in injector	✓	✗	✓	✗	✓
Water Reservoir Details					
Provided by blind holes	✓	✓	✓	✗	✓
Provided by non-smooth textures	✓	✗	✓	✗	✗
Specific water volume (0.5-3 ml)	✓	✗	✗	✗	✗

Table 2: Comparison of technical features of GPT-4o's outputs on patent EP3060168B1. ✓ means correctly stated, and ✗ means not mentioned or inconsistent with the gold claim. Detailed model outputs are in Appendix Table 10.

internal representations are compressed into a sequential, lossy verbal summary, and all subsequent generation is conditioned on this degraded intermediate rather than the original patent description.

Through detailed examination into model outputs, we conclude that the information bottleneck manifests through three concrete mechanisms.

(1) Abstraction of Critical Detail. Patent claims demand specificity: legally enforceable protection depends on concrete technical features, exact structural configurations, and precise numerical parameters—not high-level summaries. However, explicit CoT's reasoning stage systematically abstracts away the detail, which replaces specific component names and structural relationships with vague, higher-level notions. This abstraction tendency is well-suited for reasoning tasks, where generalizing from specifics is precisely what enables multi-step deduction (Wei et al., 2022). For patent claims, however, abstraction is destructive: the task requires preserving detail through generation, not reducing it through reasoning. The information bottleneck thus imposes a systematic bias toward the wrong kind of transformation, which breaks the specificity that patent claims depends on.

(2) Disruption of Internalized Generation Patterns. Through pretraining on large corpora of legal and technical documents, LLMs acquire robust internalized patterns for generating structured text—including appropriate terminology, hierarchical claim organization, and consistent cross-referencing between claim elements. Inserting an explicit reasoning phase before generation appears to disrupt these patterns. The verbalized reasoning, rather than enriching the generation process, interposes a fragmentary and incomplete conceptual plan that overrides the model's capacity to produce structurally coherent output directly from the source description.

Model	Extraction	Grouping	Formalization
Llama-3.1-8B	62	65	70
Mixtral-8×7B	70	69	73
GPT-4o	72	72	77

Table 3: Diagnostic human evaluation of intermediate steps in explicit CoT (feature extraction precision for Step 1, logical correctness for Step 2, and faithfulness for Step 3, with details in Appendix E.3).

Our findings are consistent with recent work showing that explicit CoT can be detrimental in certain tasks (Ma et al., 2025; Wang et al., 2025). For example, explicit CoT has been shown to introduce noisy and redundant reasoning traces (Zheng et al., 2025), reduce instruction-following accuracy (Li et al., 2025), and produce explanations that are misaligned with the model's internal decision process (Turpin et al., 2023). Collectively, these observations provide strong empirical support for our findings.

(3) Error Cascade across Reasoning Steps. To directly assess the quality of intermediate reasoning, we conduct a diagnostic human evaluation of each explicit CoT stage (Table 3). The results reveal a clear error cascade: Step 1 (feature extraction) achieves only 62–72% precision across models, which means essential technical features are missed. Step 2 (logical grouping) propagates and amplifies these omissions, which scores 65–72% in logical correctness. Step 3 (language formalization) generates claims conditioned on an inaccurate reasoning trajectory and imperfect adherence to the intermediate plan, which leads to compounded degradation. These errors are irrecoverable by design: once the verbalized reasoning omits a critical

feature, no subsequent step can restore it, because the model generates conditioned on the intermediate text rather than the full patent description. Implicit CoT avoids this cascade entirely by keeping intermediate representations internal, where they remain rich and uncompressed.

4.3 Case Study

Table 2 presents a detailed comparison of model outputs to support the above analysis. Explicit CoT's thinking stage abstracts "blind holes" into generic notions like "water reservoir systems" and "distribution enhancers," which loses the structural specificity essential for legal protection (Mechanism 1). It also omits structural elements such as "tray" and "side walls" from the reasoning plan, and these omissions propagate directly into the final claims (Mechanism 3). Notably, implicit CoT correctly captures both "blind holes" and the tray structure—features that the model clearly has the capacity to identify but fails to preserve when forced through the verbalization bottleneck (Mechanism 2). The full model outputs in Appendix Table 10 confirm that the explicit CoT claims follow a flawed blueprint established during reasoning, while implicit CoT produces claims with substantially higher fidelity to the gold standard.

5 Conclusion and Discussion

We investigate Chain-of-Thought prompting for patent claim generation and uncover a counter-intuitive finding: implicit CoT, where reasoning is kept internal, consistently outperforms explicit CoT, where intermediate steps are verbalized. Through diagnostic experiments, we trace this to an information bottleneck introduced by reasoning externalization—verbalized reasoning abstracts critical details, disrupts the model's internalized generation patterns, and creates error cascades across reasoning steps. We hope this work encourages more careful examination of when explicit reasoning helps or hurts, rather than treating it as a universal solution.

Limitations

This work focuses exclusively on zero-shot prompting for patent claim generation, without exploring few-shot approaches or other similar tasks, such as classification (Jiang et al., 2026a), novelty prediction (Knappich et al., 2026), and evaluation (Zuo et al., 2024; Yoo et al., 2025). Additionally, we do not perform hyperparameter tuning during inference; instead, we keep the hyperparameters fixed across all experiments to ensure fair comparisons.

Ethics Statement

Llama-3 is under *META LLAMA 3 COMMUNITY LICENSE AGREEMENT*. GPT-4o is under a commercial license provided by OpenAI, and we access it through its API. The EPD dataset is under the *CC-BY-NC-4.0* license, and the HUPD-DCG dataset is under the *CC-BY-SA-4.0* license. The use of existing artifacts is consistent with their intended use. The data does not include potential personal information or offensive content, and no ethics review board was involved.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

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*.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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, Pascal A. Scherz, and Stefan Goetz. 2025c. Towards better evaluation for generated patent claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3775–3788, Vienna, Austria. Association for Computational Linguistics.

Lekang Jiang, Wenjun Sun, and Stefan Goetz. 2026a. Reasoning for hierarchical text classification: The case of patents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11127–11142.

Lekang Jiang, Wenjun Sun, and Stephan Goetz. 2026b. Benchmarking patent drafting from inventor-style disclosures. *arXiv preprint arXiv:2608.21249*.

Lekang Jiang, Caiqi Zhang, Pascal A. Scherz, and Stefan Goetz. 2025d. 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.

Valentin Knappich, Anna Hätty, Simon Razniewski, and Annemarie Friedrich. 2026. Is it novel and why? fine-grained patent novelty prediction based on passage retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, page 845–856, New York, NY, USA. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *arXiv preprint arXiv:2505.11423*.

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.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.

OpenAI. 2024. Gpt-4o: Faster, cheaper, and more capable. `https://platform.openai.com/docs/models/gpt-4o`. Accessed: 2025-05-14.

OpenAI. 2025. Gpt-4o: Flagship gpt model for complex tasks. `https://platform.openai.com/docs/models/gpt-4.1`. Accessed: 2025-05-14.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Qiyao Wang, Guhong Chen, Hongbo Wang, Huaren Liu, Minghui Zhu, Zhifei Qin, Linwei Li, Yilin Yue, Shiqiang Wang, Jiayan Li, et al. 2025. Ipbench: Benchmarking the knowledge of large language models in intellectual property. *arXiv preprint arXiv:2504.15524*.

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, et al. 2024. Autopatent: A multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yongmin Yoo, Qiongkai Xu, and Longbing Cao. 2025. PatentScore: Multi-dimensional evaluation of LLM-generated patent claims. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30727–30746.

Yijiong Yu. 2024. Do llms really think step-by-step in implicit reasoning? *arXiv preprint arXiv:2411.15862*.

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*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

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*.

Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y Wong, and Simon See. 2025. The curse of cot: On the limitations of chain-of-thought in in-context learning. *arXiv preprint arXiv:2504.05081*.

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 Patent Background

Patent documents are important for protecting intellectual property (IP) and disclosing technical innovations. A central component of these documents is the claim set, which defines the legal scope of protection granted to the patent holder. Thus, writing high-quality claims is a critical task in patent drafting, which requires precise language to capture the essence of the invention while remaining legally enforceable. Early work investigated automated claim generation by fine-tuning language models on patent abstracts (Lee and Hsiang, 2020; Lee, 2020). However, Jiang et al. (2025d) demonstrated that this setup may be ill-posed, as abstracts tend to be generic and omit critical implementation details. Instead, they proposed the generation of claims from the full patent description, which typically contains around 10,000 tokens, compared to about 1,000 tokens in the claim set. The long contexts make the task more challenging, both in terms of content and legal requirements. Despite the difficulty, recent studies have shown that LLMs can produce high-quality claims based on the description (Jiang et al., 2025d; Wang et al., 2024; Jiang et al., 2026b). To this end, our paper explores whether zero-shot CoT prompting can further improve performance on this task.

B Experimental Details

B.1 Dataset

We conduct main experiments on the EPD dataset (Jiang et al., 2025a), which contains 693 English-language patents from the European Patent Office (EPO) for claim generation. We choose this dataset for two main reasons: (1) EPD features high-quality granted patents, in contrast to the application versions used in prior work (Jiang et al., 2025d); and (2) EPD better reflects real-world scenarios, which enables a more rigorous evaluation of LLM performance. For generalization exploration, we also test the method on the HUPD-DCG dataset (Jiang et al., 2025d), with 1,311 patents from the United States Patent and Trademark Office (USPTO), and find similar results.

B.2 Evaluation Metrics

Human Expert Evaluation: We invite qualified patent attorneys to evaluate generated claims compared to the gold claims in three aspects (scores 0–100): (1) Content Coverage: The extent to which the draft claims encapsulated all critical technical aspects of the invention. (2) Logical Structure: The degree to which the claim set exhibits a coherent, well-organized, and legally sound logical structure, including appropriate dependencies and hierarchical relationships. (3) Language Quality: The clarity, precision, and formal appropriateness of the language used in the claim set, including terminology, syntax, and adherence to patent drafting conventions. Since the evaluation process is time-intensive and costly, we randomly select 120 samples for human evaluations and each example is rated by one patent expert. Detailed instructions are introduced in Appendix C.

For automated evaluation, we use PatClaimEval (Jiang et al., 2025c), which is specifically designed for patent claim evaluation. Moreover, studies have shown that LLM-as-a-judge evaluators can achieve better human alignment, including patent claim evaluation (Liu et al., 2023; Jiang et al., 2025b).

Normal Prompt: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. Output the claims ONLY.
CoT Prompt: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. You should think step-by-step: 1. Identify the key technical features that define the invention and their importance ranking. 2. Organize these features logically based on their functional relationships and dependencies. 3. Write the claims using formal patent language to ensure clarity, consistency, and technical accuracy. Output the claims ONLY.
Explicit CoT Prompt: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. You should think step-by-step: 1. Identify the key technical features that define the invention and their importance ranking. 2. Organize these features logically based on their functional relationships and dependencies. 3. Write the claims using formal patent language to ensure clarity, consistency, and technical accuracy. Example: Input: Patent description: «Description» Output: <|beginning of thinking|> Ouput thinking <|end of thinking|> <|beginning of claims|> Output claims <|end of claims|>
CoT Prompt without Extraction: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. You should think step-by-step: 1. Organize key technical features logically based on their functional relationships and dependencies. 2. Write the claims using formal patent language to ensure clarity, consistency, and technical accuracy. Output the claims ONLY.
CoT Prompt without Logical Grouping: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. You should think step-by-step: 1. Identify the key technical features that define the invention and their importance ranking. 2. Write the claims using formal patent language to ensure clarity, consistency, and technical accuracy. Output the claims ONLY.
CoT Prompt without Formalization: You are a patent expert. Given the patent description, your task is to write high-quality patent claims. You should think step-by-step: 1. Identify the key technical features that define the invention and their importance ranking. 2. Organize key technical features logically based on their functional relationships and dependencies. Output the claims ONLY.

Table 4: Standard and CoT prompts (with variants) for patent claim generation.

Therefore, we also employ LLM-as-a-judge to evaluate claims along the same three dimensions as in human evaluation. Detailed settings and prompts are reported in Appendix D. We primarily focus on human evaluations, while treating automated metrics as supplementary evidence.

Automated Evaluation: We adopt traditional metrics for text generation, including BLEU (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and BERTScore (Zhang et al., 2019), to assess surface-level sequence overlaps and semantic similarity. In addition, we use PatClaimEval (Jiang et al., 2025c), which is specifically designed for patent claim evaluation. Moreover, studies have shown that LLM-as-a-judge evaluators can achieve better human alignment, including patent claim evaluation (Liu et al., 2023; Jiang et al., 2025b). Thus, we use the latest GPT-4.1 (OpenAI, 2025) to evaluate generated patent claims based on the above three human evaluation dimensions. Detailed settings and prompts are reported in Appendix D.

Qualitative Analysis: We provide concrete model outputs and analysis to intuitively demonstrate the performance differences among various prompting strategies. Our analysis focuses on a side-by-side comparison of claims generated from the same patent description to highlight the typical failure modes of baselines and showcase the advantages of CoT.

B.3 Environmental Settings

Detailed prompts for claim generation are reported in Table 4. The inference processes of Llama[1], Mistral[2], and Mixtral[3]are conducted on NVIDIA A100 GPUs. The total running time is about 20 hours. Inference of GPT-4o[4] is through its API. The hyperparameters we used are: temperature of 0.1, top_p of 0.95, and maximum output tokens of 4096. For BLEU, ROUGE, and BERTScore evaluations, we use the package from the HuggingFace *evaluate* library.[5]

C Human Annotations

We recruit two evaluators: a licensed patent attorney and an experienced patent apprentice who has completed formal patent drafting training and assisted in prosecuting over 50 patent applications. Both evaluators are provided with reference claims (granted versions) and candidate claims generated by different models. They are informed about the intended use of the evaluation results. The instructions are: *You are given a list of patent claims generated by different models. Compare them with the gold claim (granted version) and assess their quality. Assume the gold claim has full scores (100) of each rating aspect.*

To measure inter-annotator agreement (IAA), we compute the Quadratic Weighted Cohen's Kappa (κ) for each evaluation dimension, which is well-suited for ordinal rating scales as it penalizes larger disagreements more heavily than small ones. The results show substantial agreement: $\kappa = 0.72$ for Content Coverage, $\kappa = 0.70$ for Logical Structure, and $\kappa = 0.74$ for Language Quality. This level of agreement is sensible for this high-stakes domain. We use the scores from the licensed patent attorney as the final scores for higher reliability, and scores from the apprentice for IAA calculation.

D LLM-as-a-Judge Prompt

We use the following prompt for LLM-as-a-judge evaluations, as shown in Table 5. The hyperparameters we used are: temperature of 0.1, top_p of 0.95, and maximum output tokens of 128. To alleviate potential bias, we use GPT-4.1[6] for evaluation.

E More Results

E.1 Evaluation on More Models

We report the automated evaluation results of Mistral-7B and Mixtral-8×7B on the EPD dataset in Table 6.

E.2 Results on HUPD Dataset

We report the automated evaluation results on the HUPD-DCG dataset in Table 7.

E.3 Intermediate Step Evaluation

We report the intermediate step evaluation results in Table 8. We randomly select 80 examples for human evaluation. For feature extraction, we compare the list of features the model identifies in its thinking block against the list distilled from the gold claims. Then, we calculate the precision, recall, and F1 scores. For logical grouping, feature usage rate is measured by the number of features used in the grouping step divided by the total number of features extracted. We also rate the logical correctness in the second step, ranging from 0-100. For step three, the faithfulness score is assessed on a 100-point scale, which measures whether the final claim is a direct and logical product of the preceding reasoning. A high score indicates the final claim is a perfect synthesis of the reasoning plan, where all key elements are traceable.

E.4 Ablation Study

We conduct ablation studies to assess the effectiveness of different stages in CoT prompting. **Removing either the logical grouping or the language formalization step leads to clear performance degradation in both human and automated evaluations.** As shown in Table 9, these variants result in over 1-point average declines, with the largest drop observed in the logical structure dimension (over 1.5 points) when the logical grouping step is removed. These results underscore the critical role of both steps, especially logical grouping, in generating high-quality patent claims.

Interestingly, removing the feature extraction step slightly improves human evaluation scores. Explicitly prompting the model to extract features may fragment its focus during the more critical

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[3] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[4] GPT-4o-2024-08-06: https://platform.openai.com/docs/models/gpt-4o

[5] https://github.com/huggingface/evaluate

[6] https://openai.com/index/gpt-4-1/

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 three metrics using the referenced claims as the gold standard. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
1. Content Coverage (0–100):
The extent to which the draft claims encapsulated all critical technical 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. Logical Structure (0–100):
The degree to which the claim set exhibits a coherent, well-organized, and legally sound logical structure, including appropriate dependencies and hierarchical relationships.
0–20: Claim structure is confusing or non-compliant; dependencies are incorrect or illogical.
21–40: Some structure is evident but many dependency relationships are unclear, misplaced, or violate drafting norms.
41–60: Basic structure is acceptable; most dependencies are logical though some organizational issues persist.
61–80: Generally well-structured with minor issues in hierarchy or ordering; dependencies mostly valid.
81–100: Excellent structural clarity with fully compliant, logical, and easily navigable claim hierarchy and dependencies.

3. Language Quality (0–100):
The clarity, precision, and formal appropriateness of the language used in the claim set, including terminology, syntax, and adherence to patent drafting conventions.
0–20: Language is vague, ambiguous, or ungrammatical; terminology is inconsistent or incorrect.
21–40: Basic meaning is somewhat understandable but style is inconsistent, with notable issues in clarity or precision.
41–60: Mostly clear and accurate, though with occasional awkward phrasing or imprecise terminology.
61–80: Generally polished and professional language with only minor lapses in clarity or consistency.
81–100: Language is highly precise, concise, and consistent, following patent norms and clearly conveying the technical content.

Evaluation Steps:
1. Read the referenced claims carefully and assume the referenced claims have scores of 100 in all Evaluation Criteria.
2. Read the draft claims and compare it 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):
- Content Coverage: X,
- Logical Structure: X,
- Language Quality: X.

Table 5: LLM-as-a-judge prompt used for claim evaluation.

	Traditional Metrics			PatClaimEval	LLM-as-a-Judge			
Model	**BLEU**	**R-L**	**BERTScore**	**Quality**	**Content**	**Logic**	**Language**	**Average**
Mistral-7B	31.4	41.3	85.6	17.6	72.8	68.9	68.9	69.9
w/ Implicit CoT	31.0	40.1	85.5	18.1	73.4	69.5	68.3	70.1
w/ Explicit CoT	15.5	33.0	83.9	16.3	63.2	60.3	56.8	59.9
Llama-3.1-70B	27.2	38.9	85.6	20.9	75.9	74.2	74.3	74.5
w/ Implicit CoT	26.3	38.2	85.6	21.1	75.5	75.0	74.3	74.5
w/ Explicit CoT	11.7	35.9	84.7	22.0	61.0	64.0	65.0	63.1

Table 6: Automated evaluation results of different models on EPD dataset.

logical grouping phase. Since logical grouping inherently involves identifying and organizing key features, an additional extraction step can introduce unnecessary overhead and distract the model from fully engaging its reasoning and planning abilities. Removing this step allows the model to implicitly perform feature identification while fully focusing on logical grouping to improve overall

Model	Traditional Metrics			PatClaimEval	LLM-as-a-Judge			
	BLEU	R-L	BERTScore	Quality	Content	Logic	Language	Average
Mistral-7B	36.0	40.6	87.0	16.5	71.4	71.5	69.4	70.5
w/ Implicit CoT	35.6	40.9	86.9	18.4	72.3	72.9	70.9	71.7
w/ Explicit CoT	19.6	33.0	85.1	18.3	62.7	62.2	60.3	61.5
GPT-4o	22.3	39.0	86.5	22.0	73.2	73.3	72.2	72.6
w/ Implicit CoT	20.4	37.6	86.4	23.0	73.4	74.1	73.3	73.4
w/ Explicit CoT	13.5	33.0	86.0	22.4	65.8	71.0	71.3	69.0

Table 7: Automated evaluation results of different models on the HPUD-DCG dataset.

Model	Step 1: Extraction			Step 2: Grouping		Step 3: Formalization
	Precision	Recall	F1 Scores	Usage	Correctness	Faithfulness
Llama-3.1-8B	62	56	58	98	65	70
Mixtral-8×7B	70	53	60	99	69	73
GPT-4o	72	58	64	99	72	77

Table 8: Human analysis on intermediate steps of explicit CoT prompting.

Model / Setting	PatClaimEval	LLM-as-a-Judge				Human Evaluation			
	Quality	Content	Logic	Language	Average	Content	Logic	Language	Average
GPT-4o	23.5	77.3	76.4	74.8	75.9	88.5	86.3	87.7	87.5
w/ Implicit CoT	23.7	78.0	76.8	75.4	76.5	93.0	92.2	92.8	92.7
w/o Extraction	23.5	78.1	76.4	75.0	76.2	93.3	92.7	93.2	93.1
w/o Grouping	23.4	77.0	75.5	74.4	75.4	91.7	90.7	92.2	91.5
w/o Formalization	24.3	77.2	75.0	74.1	75.2	91.3	91.8	91.8	91.7

Table 9: Automated and human evaluation results of ablation study on GPT-4o.

performance.

In summary, our findings suggest that logical grouping is the most critical stage. Simplifying the CoT prompt to focus on this important stage, which aligns closely with the LLMs' strengths in reasoning and organization, can further enhance claim generation performance.

Gold claims 1. A combination (10) comprising: a) an air tight container (12) having a water reservoir (14); and b) an injector (60) being retained in the container (12), wherein the injector is adapted to fold and inject an intraocular lens (16) into an eye of a patient; and c) the intraocular lens (16) being foldable and hydrophilic and disposed unfolded within the injector (60) so as not to be in direct contact with the water reservoir (14), characterized in that the container (12) comprises a tray (18) having an elongate compartment (20) wherein the tray (18) comprises a bottom wall (22) and side walls (26) wherein the water reservoir (14) is provided by blind holes (54) in interior surfaces of the bottom wall (22) and side walls (26) of the tray (18), wherein the water reservoir (14) retains between 0.5 and 3 ml of water, wherein the blind holes span a majority of the length and width of the tray (18). 2. The combination (10) of claim 1 wherein the water reservoir (14) is provided by non-smooth textures (56) on the interior surfaces of the bottom wall (22) and side walls (26) of the tray (18) and wherein the water reservoir (14) comprises water retained within the surfaces with non- smooth textures (56).

GPT-4o (standard prompt) 1. A packaging system for a foldable intraocular lens (IOL), comprising: a. a substantially air-tight container having an interior volume; b. a water reservoir configured to maintain a humidity level of at least about 90% within the container; c. an intraocular lens disposed within the container, wherein the lens is not immersed in liquid but is exposed to the humidity maintained by the water reservoir; d. a cover sealing the container to minimize water loss through diffusion; e. wherein the container is configured to allow condensation on the lens surfaces to maintain hydration adequate for foldability and delivery through an injection tube. 2. The packaging system of claim 1, wherein the water reservoir comprises: a. a water permeable pouch disposed within the container, the pouch containing water that is released over time to maintain the desired humidity level. 3. The packaging system of claim 1, wherein the water reservoir comprises: a. a water absorbing polymer capable of releasing water to the atmosphere within the container. 4. The packaging system of claim 1, wherein the container includes: a. void volumes separate from the location of the intraocular lens, configured to retain free water and minimize humidity gradients. 5. The packaging system of claim 1, further comprising: a. distribution enhancers within the container to facilitate even distribution of water vapor, selected from the group consisting of towels with wicking properties, pieces of plastic with blind holes, and textured surfaces on the interior of the container walls. 6. The packaging system of claim 1, wherein the container is autoclavable and capable of withstanding temperatures of at least 121 degrees Celsius for sterilization purposes. 7. The packaging system of claim 1, further comprising: a. an injector disposed within the container, adapted to fold and inject the intraocular lens through a cylindrical injection tube having an inside diameter of 2 millimeters or less. 8. The packaging system of claim 1, further comprising: a. an autoclavable foil pouch enclosing the container, providing a second sterile barrier and minimizing water loss through diffusion.

GPT-4o (Implicit CoT) 1. A packaging system for a foldable intraocular lens (IOL), comprising: a. a substantially air-tight container having an interior volume, said container comprising a tray with a bottom wall, side walls, and an elongate top opening; b. a cover sealing the elongate top opening of the container, said cover comprising a multilayered foil lid to minimize water loss through diffusion; c. a water reservoir within the container, said reservoir providing an amount of water sufficient to moisture saturate the interior volume of the container at temperatures up to those required for steam sterilization, and to form small water droplets throughout the container; d. wherein the water reservoir comprises a water permeable pouch disposed below the cover, said pouch designed to capture and retain water, releasing it as free water when atmospheric and other free water within the container is depleted over time by diffusion; e. wherein the container is configured to maintain a humidity level of at least about 90% within the interior volume to ensure adequate hydration of the IOL for foldability and insertion into an eye. 2. The packaging system of claim 1, wherein the water reservoir further comprises distribution enhancers selected from the group consisting of: a. a towel with wicking or sponge-like properties; b. a piece of plastic with openings in the form of blind holes that span the surfaces of the tray; c. a water absorbing polymer having a high water content that can be released to the atmosphere of the container, said polymer defining holes in which free water can reside. 3. The packaging system of claim 1, wherein the tray is made from a thermoplastic material, such as polypropylene, and includes a circumferential rim with a raised outer lip surrounding a recessed inner band. 4. The packaging system of claim 1, wherein the water reservoir is further provided by void volumes within the container, said void volumes being defined separate from the location of the intraocular lens within the container, and wherein the void volumes are maximized to minimize humidity gradients within the container. 5. The packaging system of claim 1, wherein the water reservoir is further provided by blind holes in interior surfaces of the tray walls, said blind holes spanning a majority of the length and width of the tray, and/or by a non-smooth texture on the interior surfaces of the tray walls capable of retaining small amounts of free water. 6. The packaging system of claim 1, wherein the combination of the container and water reservoir is autoclavable, capable of withstanding heating to 121 degrees Celsius or more for a period of a half hour or more. 7. The packaging system of claim 1, further comprising an autoclavable foil pouch enclosing the container, said pouch minimizing water loss to diffusion, providing a second sterile barrier, and allowing the container to be introduced into a sterile field. 8. The packaging system of claim 1, wherein the intraocular lens is disposed unfolded within an injector, said injector adapted to fold and inject the intraocular lens into the eye of a patient through a cylindrical injection tube having an inside diameter of 2 millimeters or less.

Continue on next page

GPT-4o (Explicit CoT) <|beginning of thinking|> 1. Key Technical Features and Importance Ranking: - Packaging system for foldable intraocular lenses (IOLs) using water-saturated or nearly saturated air instead of immersion in liquid (High importance). - Substantially air-tight container with minimal air volume to maintain hydration (High importance). - Water reservoir system that maintains humidity and prevents water loss (Medium importance). - Distribution enhancers for even water distribution within the container (Medium importance). - Autoclavable packaging for sterilization (Medium importance). - Use of a foil pouch for additional sealing and sterile barrier (Low importance). 2. Logical Organization: - Begin with the overall packaging system and its purpose. - Describe the container's structure and its air-tight properties. - Explain the water reservoir system and its function. - Detail the distribution enhancers and their role in maintaining humidity. - Include the autoclavable feature for sterilization. - Conclude with the additional sealing provided by the foil pouch. 3. Writing the Claims: - Use formal patent language to clearly define each feature and its function. - Ensure each claim is independent and builds logically on the previous one. - Maintain clarity and technical accuracy throughout the claims. <|end of thinking|> <|beginning of claims|> 1. A packaging system for foldable intraocular lenses, comprising: a substantially air-tight container configured to house an intraocular lens, wherein the container maintains water-saturated or nearly saturated air to preserve the hydration of the lens without immersing it in liquid. 2. The packaging system of claim 1, wherein the container comprises: a tray with an elongate compartment bounded by a bottom wall, side walls, and an elongate top opening, the top opening being sealed by a multilayered foil lid to minimize water loss through diffusion. 3. The packaging system of claim 2, further comprising: a water reservoir within the container, the reservoir providing an amount of water sufficient to moisture saturate the interior volume of the container and form small water droplets throughout the container. 4. The packaging system of claim 3, wherein the water reservoir includes: a water permeable pouch disposed within the container, the pouch designed to capture and retain water, releasing it as needed to maintain humidity levels. 5. The packaging system of claim 4, further comprising: distribution enhancers within the container to facilitate even distribution of water, the enhancers including materials with wicking properties or structures with blind holes to retain free water. 6. The packaging system of claim 1, wherein the container is autoclavable, capable of withstanding temperatures of at least 121 degrees Celsius for sterilization purposes. 7. The packaging system of claim 1, further comprising: an autoclavable foil pouch enclosing the container, providing an additional sterile barrier and minimizing water loss through diffusion. 8. The packaging system of claim 1, wherein the intraocular lens is disposed unfolded within an injector, the injector adapted to fold and inject the lens through a cylindrical injection tube with an inside diameter of 2 millimeters or less. <|end of claims|>

Table 10: Example model outputs of patent EP3060168B1.