Distilling Examples into Task Instructions: Enhanced In-Context Learning for Real-World B2B Conversations

Anonymous ACL submission

Abstract

In-context learning (ICL) is the standard method for low-data classification, yet its efficacy in specialized, intricate domains remains largely unexplored. We address the challenge of classifying semantically complex, multi-party B2B conversations, where traditional ICL encounters significant limitations, especially as context length increases due to concatenation of multiple few-shot examples. We introduce the `Call Playbook` dataset, featuring five classification tasks derived from real-world B2B conversations targeting core sales concepts. To bridge the gap between performance and practical utility, we propose novel knowledge extraction methods that distill verbose examples into compact, interpretable representations of structured classification criteria and precise task descriptions. Our approach achieves a 99% reduction in token usage and improves macro-averaged AUC by up to 7% over traditional ICL. Notably, our method remains robust as context grows, unlike advanced token compression baselines which degrade by over 9 points. Our interpretable artifacts facilitate seamless refinement, allowing users to directly modify classification logic. This approach addresses critical needs for transparency, efficiency, and user interaction in real-world NLP applications.[1]

1 Introduction

Recent advances in large language models (LLMs) have transformed natural language processing (NLP), particularly through in-context learning (ICL), where models perform tasks by conditioning on few examples without parameter updates (Brown et al., 2020; Min et al., 2022). This paradigm, commonly implemented through few-shot learning approaches, has demonstrated remarkable success across diverse NLP tasks (Sanh et al., 2022; Chowdhery et al., 2023). However, its application to specialized domains such as business-to-business (B2B) sales communications presents unique challenges that have yet to be thoroughly examined (Gupta et al., 2022; Chamieh et al., 2024).

In B2B sales environments, analyzing prospect conversations is essential to extract actionable intelligence guiding deal strategy (Grosz et al., 1995; Dean et al., 2017). To achieve this at scale, organizations must automatically classify conversational segments across diverse and evolving task intents that are not known in advance. These tasks operate under severe constraints: limited labeled data, minimal annotation overhead, and the impracticality of fine-tuning models for each emerging intent.

While ICL is theoretically suited to these constraints, it faces significant hurdles in practice. Notably, while current LLMs can potentially support up to millions of tokens in their context windows (Reid et al., 2024), our analysis shows that these models exhibit severe performance degradation under standard ICL with far fewer tokens in practice. This degradation stems mainly from the concatenation of multiple few-shot examples, challenging the conventional assumption that more examples improve performance (Agarwal et al., 2024; Bertsch et al., 2024). Furthermore, B2B classification requires high levels of interpretability and transparency for business professionals to validate decisions and maintain trust (Doshi-Velez and Kim, 2017; Lipton, 2018).

To provide a basis for evaluation, we first introduce five novel B2B classification datasets derived from real-world sales conversations. These datasets target prospect understanding across fundamental sales concepts and serve to highlight the limitations of standard ICL in high-stakes business contexts.

We then present a framework designed to overcome these limitations by distilling classification knowledge into compact and interpretable formats. Unlike token-level compression methods that yield unreadable, fragmented tokens (Li et al., 2023) or computationally expensive approaches (Xu et al.,

[1] Data and code will be released upon acceptance.

2024), our method automatically extracts coherent task knowledge from few-shot examples and converts them into explicit classification criteria. These transparent artifacts facilitate both automated processing and human oversight, making our approach uniquely suited for the efficiency and intent flexibility required in practical B2B applications.

Our contributions are multifold: (1) We introduce a new B2B dataset spanning five core sales concepts. (2) We propose novel ICL methods that shift from example concatenation to automated knowledge distillation. (3) Extensive experiments demonstrate superior performance over standard few-shot baselines and advanced token compression methods. (4) Our approach yields significant computational savings through token reduction and faster inference times. (5) The framework offers interpretability, enabling human-in-the-loop enhancement and user guidance.

2 Related Work

2.1 In-Context Learning (ICL)

In-context learning (ICL) has emerged as a major development in NLP, showcasing the ability of LLMs to perform diverse tasks using a few labeled samples provided in the prompt, without requiring explicit fine-tuning (Brown et al., 2020; Dong et al., 2024). This capability, first demonstrated by models such as GPT-3, has spurred extensive research into its mechanisms, performance, and limitations.

Much of this research explores the role of demonstrations in ICL, focusing on prompt design and the selection and ordering of examples (Liu et al., 2022; Pan et al., 2023). Strategies include retrieving semantically similar instances (Rubin et al., 2022) or modeling inter-example relationships (Ye et al., 2023), as well as employing active learning to select effective subsets (Zhang et al., 2022). In contrast, this work derives task-specific knowledge directly from a small number of examples.

Recent studies have shown that ICL struggles with long contexts, especially when many examples are used (Li et al., 2024; Lee et al., 2025; Modarressi et al., 2025). To address this challenge, compression techniques have been explored for reducing example length in NLP applications. Token-level methods remove redundant content based on self-information metrics (Li et al., 2023) or perform iterative compression through instruction-tuned models (Jiang et al., 2023b; Pan et al., 2024), but are limited to selecting subsets of existing tokens, often resulting in incoherent examples.

Model-based approaches train dedicated compression models for specific contexts, such as abstractive summarization for retrieval-augmented generation (Xu et al., 2024) or active compression for question answering (Yoon et al., 2024). While effective, these methods require labeled datasets and specialized model training. Unlike these approaches, our method generates concise, interpretable demonstrations that preserve coherence while distilling essential task knowledge without requiring additional model training.

2.2 NLP for the B2B Domain

B2B conversations pose distinct challenges for NLP due to their complexity, involving multiple stakeholders, long sales cycles, and domain-specific language (Grewal et al., 2022; Wu et al., 2024). Business signals are often implicit (Voria et al., 2024), positioning B2B conversations as a rigorous benchmark for evaluating LLM ability to interpret nuanced, context-specific language.

NLP is increasingly applied in the B2B domain for business goal identification (Campbell et al., 2003; Spruit et al., 2021), sales enhancement (Patel et al., 2022), understanding sales conversations (Chai et al., 2001), customer segmentation (Lieder et al., 2019), and sales forecasting (Bohanec et al., 2017; Zahid et al., 2021; Rohaan et al., 2022). However, existing work lacks consideration of the fluid multi-stakeholder dynamics and evolving context throughout deal progression.

To bridge this gap, this paper presents a novel dataset of authentic B2B sales conversations from real-world interactions with multiple stakeholders across different deal stages, annotated for diverse interconnected prospect-focused business concepts. Unlike existing B2B datasets that typically focus on customer support tickets, product reviews, or short transactional exchanges, our dataset captures extended multi-party sales dialogues that reflect the complexity of actual business negotiations.

In this work, we demonstrate how LLMs can automatically generate user-guiding task instructions to enhance ICL, enabling comprehensive analysis of prospect communication patterns throughout these extended dialogues.

3 The Call Playbook Dataset

This section describes the data collection, text processing, and annotation processes used to construct

the `Call Playbook` dataset.

3.1 Dataset Overview

We constructed the `Call Playbook` dataset from 50 English B2B sales calls ranging from 30 to 90 minutes. Each transcript is structured as a sequence of monologues attributed to speakers on either the seller or the prospect side. The dataset contains annotations for five key sales concepts targeting the prospect side. Each concept captures a distinct aspect of the prospect's intent and purchasing process:

- **Business Goals** describe the outcomes or objectives the prospect wishes to achieve.
- **Decision Criteria** refer to the standards used to evaluate potential solutions.
- **Decision Makers** identify individuals or roles involved in making the purchasing decision.
- **Decision Making Process** refers to the sequence of steps the prospects follow when making a decision.
- **Pain Points** reflect the challenges and obstacles the prospect seeks to address.

3.2 Data Annotation

The annotation task involved three trained in-house annotators who identified and labeled textual spans where each concept was expressed, guided by a domain expert who supervised the process. Each call was segmented into overlapping *snippets* of five consecutive monologues with an overlap of one monologue between adjacent snippets.[2] This segmentation produces dense, contextually grounded segments that preserve the flow of conversation while enabling targeted classification. Each snippet was then assigned binary labels for the five concepts: a snippet was marked *positive* if it contained at least one annotated span for a concept, and *negative* otherwise. Any disagreements that occurred between annotators were resolved through discussion until consensus was reached.

For each concept, we created class-balanced train and test sets containing 200 samples each. When positive examples were limited, we evenly split them between sets and filled the remainder with randomly sampled negatives. To avoid data leakage, we ensured calls do not overlap between sets. This balanced construction establishes a rigorous benchmark that prevents evaluation metrics from being masked by majority-class prevalence.

[2]Short monologues containing fewer than five words are excluded from the count.

Data curation procedures and usage guidelines are outlined in Appendix A. The data set was thoroughly processed and anonymized to protect sensitive information, as detailed in Appendix B. Key dataset statistics and representative examples, illustrating its diversity and complexity, are provided in Appendix C.

4 Methodology

This section presents our approach to enhancing the effectiveness of ICL for classification tasks in B2B conversational analysis. We begin by formalizing the problem setup and then describe our novel methods for improving few-shot classification performance within this framework.

4.1 Problem Setup

To reflect practical constraints in conversational analysis, we consider the problem of classifying conversational segments into domain-relevant categories under conditions of minimal task specifications and few labeled examples.

We define this setup formally as follows: Given a short user-provided task intent i and a labeled dataset $E = (x_j, y_j)_{j=1}^{M}$, where each x_j is a conversational segment and $y_j \in C$ is its corresponding class label from the set of possible classes C, our goal is to build a classifier $f : X \to C$ that can accurately predict labels for new conversation snippets. In our experimental setup, we focus on binary classification where $C \in \{0, 1\}$, with 1 indicating the presence of the target concept and 0 indicating its absence.

This classification task presents multiple challenges that make ICL particularly suitable for practical B2B applications: (1) **Limited data availability**: specialized domains often have scarce labeled examples due to their specific nature and privacy constraints; (2) **Minimal user overhead**: practitioners need quick deployment without extensive annotation or model training; (3) **Task diversity**: user intents vary widely, making task-specific fine-tuning of multiple models impractical at scale.

ICL addresses these constraints by leveraging a general LLM to classify diverse task intents using only a few examples; our framework extends this capability by extracting interpretable knowledge that enables human-in-the-loop intervention.

4.2 Classification Process

Our general classification framework, illustrated in Figure 1, proceeds as follows:

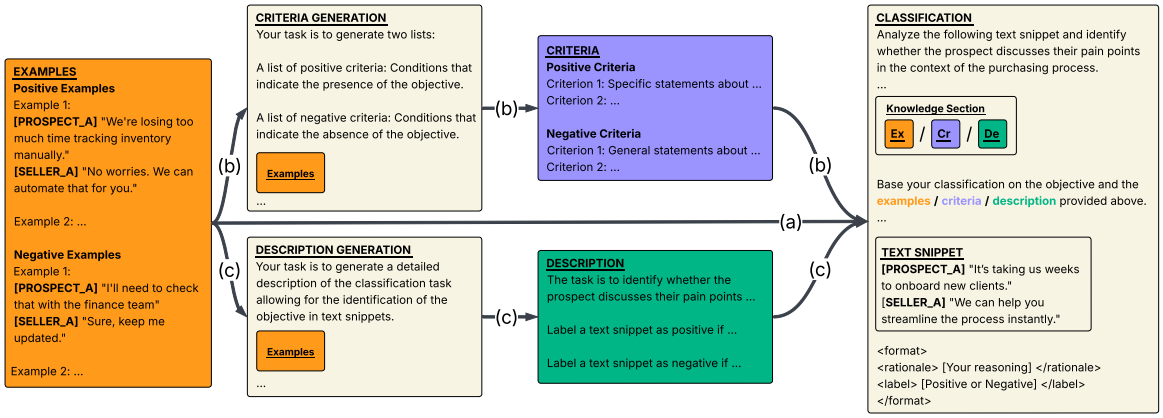

Figure 1: Classification framework overview. The process begins by sampling labeled examples (a), which are either used directly in the knowledge section (traditional few-shot learning) or transformed through knowledge extraction into criteria lists (b) or a task description (c). This extracted knowledge replaces raw examples in the classification prompt, producing structured predictions with explicit rationales for enhanced classification performance.

Step 1: Sample Selection. We randomly sample a small subset of N labeled examples from the training set E, preserving the original class distribution, to serve as the basis for ICL. This step is shared across all proposed methods.

Step 2: Knowledge Extraction. Based on these examples, we either use them directly in a standard few-shot prompt or transform them into distilled knowledge, such as criteria or detailed descriptions, using an LLM. When employed, we use the same LLM for this step and for the classification task described in Step 4 for consistency.

Step 3: Prompt Construction. We construct a classification prompt that includes the user intent i, either the sampled examples or the derived knowledge, and the test snippet x_{test} to be classified. For all methods, we use the same classification prompt template, further discussed in Appendix D.

Step 4: Classification. We query an LLM with the constructed prompt to obtain a classification prediction $\hat{y} \in C$ for each test snippet.

Rather than relying on raw examples in the prompt, our method introduces a knowledge extraction phase that summarizes them into generalizable knowledge. This not only overcomes token limitations in few-shot setups but also supports strong performance and improved user interpretability.

4.3 Classification Methods

4.3.1 Few-shot Learning (Examples)

The standard few-shot learning approach, which we denote as "Examples", serves as our main baseline. In this method, we directly include the N sampled examples in the classification prompt to provide guidance for the model, as shown in Figure 6.

While this approach has proven effective for many tasks, it faces notable challenges in conversational domains. Each snippet can be hundreds of words long (see Table 3), quickly consuming token budget as N increases. Also, the model must infer classification rules implicitly from examples, which may lead to poor generalization, especially when the task is difficult or underspecified.

4.3.2 Summary-Ex Method

To harness the advantages of information compression and reduced contextual noise, we introduce "Summary-Ex" (Summary from **Ex**amples) as an intermediate baseline. This method replaces the full examples with concise summaries that retain essential discriminative information:

Step 2: Example Summarization. We prompt an LLM to generate a brief summary for each example, condensing it to 3-5 sentences while preserving the original conversation format and speaker affiliations. The summarization process focuses on removing redundant information and filler words while retaining all discussed business content and maintaining the essential structure and flow of the conversation. The detailed prompt for summarization is provided in Appendix E.1.

Steps 3-4: Classification with Summarized Examples. We substitute the original sampled examples with their summarized variants within the standard few-shot prompt format to classify the original text snippets.

This approach reduces token usage compared to full examples while maintaining the intuitive example-based learning paradigm. However, it still requires the model to implicitly infer classification patterns and may lose important contextual nuances during the summarization process.

4.3.3 Criteria-Ex Method

Our first alternative, "Criteria-Ex" (Criteria from **Ex**amples), extracts explicit classification criteria from training examples rather than relying on implicit pattern inference:

Step 2: Knowledge Extraction. We instruct the relevant LLM to generate two lists of criteria based on the sampled examples and the user intent i: (1) a list of positive criteria indicating the presence of the concept in the text snippet, and (2) a list of negative criteria indicating its absence. The prompt (detailed in Appendix E.2) directs the model to analyze the distinguishing patterns between positive and negative examples. For binary B2B tasks, this entails identifying patterns that indicate whether a concept is discussed or not in the conversation.

Steps 3-4: Classification with Criteria. We proceed with classification by replacing the few-shot examples in the classification prompt with the generated criteria (see Figure 7).

This approach significantly reduces token usage compared to few-shot examples, as the criteria typically require far fewer tokens than the original conversation snippets. It also enhances explainability by making classification logic explicit rather than implicit, and improves generalization by extracting patterns rather than relying on specific examples.

4.3.4 Description-Ex Method

Our next method, "Description-Ex", similarly transforms **Ex**amples into a detailed task description that extends beyond the short user intent:

Step 2: Knowledge Extraction. We prompt the relevant LLM to analyze the sampled examples and user intent i to generate a comprehensive task description that captures the essence of the classification task. The description explains the characteristics that indicate the presence or absence of the target concept, and provides a coherent explanation of the concept boundaries. The prompt for generating descriptions is provided in Appendix E.3.

Steps 3-4: Classification with Description. We replace the few-shot examples with the generated description in our classification prompt (see Figure 8) and proceed with classification.

This approach offers benefits similar to Criteria-Ex but provides a cohesive narrative explanation that may better capture complex relationships and context-dependent aspects of classification. Its structured format aligns more naturally with how LLMs process instructions, which can improve generalization in challenging scenarios where rigid criteria may overlook important subtleties.

4.3.5 Iterative Improvement Methods

We further investigate whether our proposed methods can be enhanced through iteration, examining the potential for refining derived knowledge.

Criteria-De generates classification criteria from a previously generated task **De**scription, produced by Description-Ex, rather than directly from examples. The associated prompt is detailed in Appendix E.2.

Description-Cr generates a task description from previously generated **Cr**iteria, produced by Criteria-Ex, testing whether structured criteria can be expanded into a more comprehensive narrative. The associated prompt is detailed in Appendix E.3.

These iterative variants illustrate how knowledge representations can be progressively refined, e.g., from examples to descriptions to criteria and back, revealing their complementary roles. This capacity for stepwise enhancement renders the approach especially well-suited for dynamic, human-in-the-loop workflows, where evolving guidance or labeling needs are met through successive refinements rather than redesigning prompts from scratch.

5 Experiments

We evaluate all methods on the five concepts of `Call Playbook`. Using ICL, we systematically vary the number of examples (0, 10, 25, 50, 75, and 100) to assess performance across different few-shot levels[3].

Examples are randomly sampled while preserving the original class distribution. To mitigate potential biases and ensure statistical reliability, we repeat each configuration five times with different random samples and report averaged metrics.

Our evaluation covers five LLMs of varying capabilities: GPT-4o (Hurst et al., 2024), Claude Sonnet 3.7 and Claude Haiku 3 (Anthropic, 2024), as well as Mistral Large and Mistral Small (Jiang et al., 2023a). This selection includes both proprietary

[3] In zero-shot scenarios, our criteria- and description-based methods distill information solely from the user intent, leveraging general knowledge to enable operation without examples.

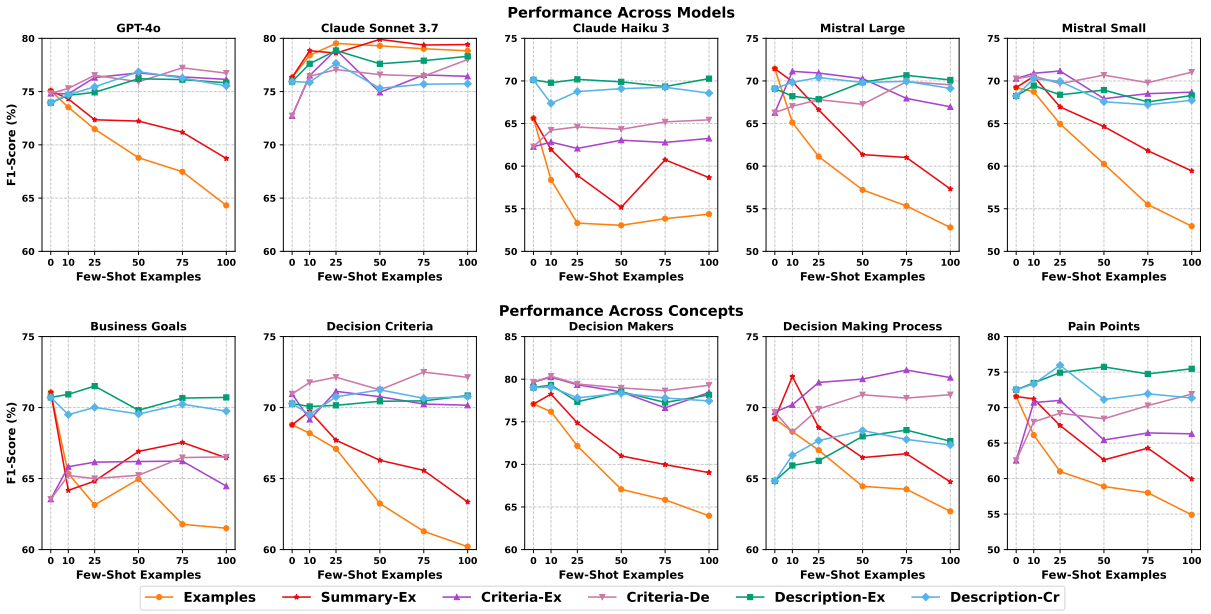

Figure 2: Top: Macro-average F1 performance for each model, averaged over all concepts. Bottom: Macro-average F1 performance for each concept, averaged over all models.

and open-weight models with diverse architectures and parameter scales. To ensure a fair comparison, we use the same set of examples across all models and set *temperature* to 0 for deterministic outputs.

For implementation, we use the *LangChain* framework[4] with the Azure API for GPT-4o and the AWS Bedrock API for the remaining models.

6 Results

6.1 Main Results

Our ICL analysis across various LLMs and datasets reveals key classification performance patterns. Figure 2 presents a comprehensive view of macro-average F1 scores, with the top row showing average performance across models and the bottom row displaying results across concepts. The complete results for all models and concepts are presented in Appendix L.1. Detailed qualitative analyses of our methods are provided in Appendices F, G, H, and I.

Method Performance Overview Across all experiments, **our advanced prompting methods consistently outperform the basic example-based methods**, with criteria and description variants performing similarly. AUC analysis shows a tight cluster: Criteria (De: 76.3%, Ex: 76.2%), Description (Cr: 75.9%, Ex: 75.6%), while Summary-Ex (72%) and Examples (69.3%) trail behind.

[4] https://www.langchain.com/

Few-shot Scaling Patterns Our analysis reveals distinct patterns in average macro-F1 scores as few-shot examples increase. The most striking finding is that **the standard few-shot (Examples) method exhibits severe performance degradation as the number of few-shot examples increases**. This phenomenon is consistent across all experiments, highlighting the inherent difficulty of our datasets. The Examples method degrades dramatically from 71.5% at 0-shot to 60.7% at 100-shot on average. This observation is also true for the Summary-Ex method, which shows similar degradation down to 64.7%, implying that summarization alone cannot fully address this challenge.

This declining performance aligns with recent findings on long-context models (Li et al., 2024; Lee et al., 2025; Modarressi et al., 2025), which demonstrate that performance often plateaus or degrades as context size increases. Our B2B dataset's complex relational structures and specialized nuances seem to exacerbate this degradation, causing traditional ICL to struggle with more examples.

Remarkably, our proposed methods show robustness to context length, with **Description-Ex showing steady improvement** from 0-shot (71.5%) to 100-shot (72.6%) with minimal fluctuation. **Criteria-De demonstrates the largest overall gains** from 69.3% to 72.2%. In contrast, **Description-Cr and Criteria-Ex peak at intermediate shot counts**, both at 25 shots with 72.4%

Method / # Shots	0	10	25	50	75	100	Avg
Criteria-Ex	69.3	71.2	71.9	70.6	70.4	70.3	70.6
Description-Ex	**71.5**	**71.9**	**72.0**	**72.5**	**72.3**	**72.6**	**72.1**
SC	71.4	69.8	67.4	64.7	63.8	62.1	66.5
LLMLingua-2	46.4	49.1	48.9	49.0	46.8	47.6	48.0

Table 1: Comparison with token compression methods across few-shot example counts. Average macro-average F1 scores over all models and datasets.

and 71.9% respectively, before declining to 71.3% and 70.3%, indicating performance saturation.

Model Comparison Figure 2 (Top) indicates that **Sonnet 3.7 achieves the highest performance overall** with 77% macro-average F1, showing strong scaling and peak performance at 25 shots. GPT-4o performs consistently until 50 shots before declining, while both Mistral models exhibit clear degradation as shots increase, with Mistral Small peaking early at 10 shots. Haiku 3 shows minimal variation between medium and high shots after an initial performance decline. These findings suggest that **larger models often benefit from additional exemplars, while smaller models struggle with increased context length.**

Concept Patterns The performance across concepts, shown in Figure 2 (Bottom), reveals distinct patterns in classification effectiveness and method suitability. Decision Makers consistently achieves the highest F1 scores (reaching 80%) across all methods, indicating that identifying stakeholders involves more recognizable linguistic patterns than other concepts. Notably, Business Goals and Decision Criteria show similar performance profiles with modest differences between methods, suggesting these concepts share comparable semantic structures in B2B conversations. In contrast, Pain Points exhibits the widest performance spread (55%–75% F1) with description-based methods decisively outperforming others, confirming that descriptive context significantly enhances the model's ability to recognize problem-oriented language. Finally, the procedural concept of Decision Making Process uniquely favors Criteria-Ex, showing consistent improvement as shot count increases.

These results demonstrate that **concept characteristics determine optimal prompting strategies: abstract concepts benefit from descriptive context while systematic concepts require structured criteria.** This underscores the need for concept-specific approaches in B2B sales analysis.

6.2 Token-Level Compression Comparison

To isolate whether our performance gains stem from distillation quality or merely token reduction, we compare against two extractive baselines that prune original tokens without generating new content: LLMLingua-2 (Pan et al., 2024), which compresses entire prompts, and Selective Context (SC) (Li et al., 2023), which targets example compression specifically. We apply a standard compression rate of 50% to balance aggressive compression with information retention. Table 1 presents results averaged across all models and datasets, with detailed per-dataset breakdowns in Appendix L.2.

Consistent with our earlier observations, our methods show improving or stable performance as examples increase, while compression baselines show opposite trends. LLMLingua-2 remains below 50% throughout, while SC declines substantially from 71.4% to 62.1% (a 9.3 point drop), negating the benefit of additional examples.

These divergent trajectories reveal fundamental limitations of these methods. **Extractive token compression either prunes essential task instructions or yields fragmented examples composed of disconnected salient tokens, failing to provide coherent guidance as example density increases.** LLMLingua-2 fails because it cannot distinguish between compressible content and task-defining elements. SC avoids this by targeting examples specifically. However, its independent token-level pruning produces compressed examples that lack interpretable structure and relational connectivity. Consequently, as more examples are added, this lack of internal coherence fails to improve performance. Moreover, like traditional prompting (Section 6.4), both methods scale linearly with example count and require training specialized models, introducing significant preprocessing overhead.

In contrast, our methods distill knowledge into interpretable, generalizable patterns through structured distillation. This yields both superior performance and human-readable representations via a single LLM call that does not require additional training.

6.3 Human-in-the-Loop Enhancement

To assess the interpretability and extent of human contribution to our methods, we conducted an experiment involving human annotators. We selected the criteria and descriptions generated by Sonnet 3.7 (our strongest model) from the most effective

Model / Annotator	Criteria-Ex	Description-Ex
Claude Sonnet 3.7	77.94	79.84
Annotator 1	**80.59**	**81.84**
Annotator 2	75.63	79.80
Annotator 3	79.52	80.90
Annotator 4	80.30	80.50
Annotator 5	77.92	80.81

Table 2: Average macro-average F1 comparing model-generated knowledge against human-refined versions.

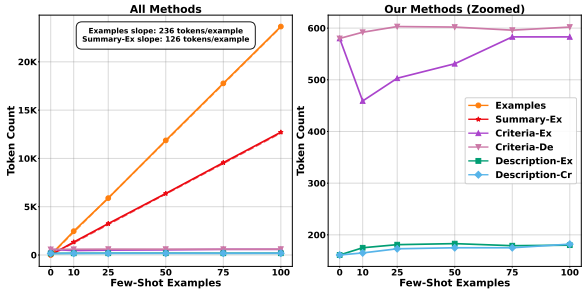

Figure 3: Token count analysis. Left: All methods, showing linear growth for the Examples and Summary methods, while the Criteria and Description methods remain flat. Right: Zoomed view of our methods, revealing minimal variation as the example count increases.

few-shot size of 25 examples from our first iteration. Five human annotators were then asked to modify the texts generated by Criteria-Ex and Description-Ex for all five concepts. Annotators were given full freedom to revise the generated text by editing, removing, or adding content, while preserving reasonable similarity to the original output.

Table 2 compares the original model-generated elements and the human-modified versions in terms of average macro-average F1 scores across the five tasks. Notably, three annotators improved Criteria-Ex performance and one achieved similar results, while four improved Description-Ex performance, with gains up to 2.65% and 2%, respectively. Descriptions proved more conducive to human refinement than criteria, likely because descriptions allow for more flexible reformulation without altering the underlying logic, while criteria modifications can more easily introduce unintended logical conflicts or overlook boundary conditions. Unlike traditional few-shot approaches where classification logic remains opaque, **our methods provide interpretable artifacts that enable effective human-in-the-loop collaboration, successfully combining LLM strengths with human expertise.** A detailed analysis of human modifications and their impact on performance is provided in Appendix J.

6.4 Computational Efficiency Analysis

Figure 3 compares token consumption across methods using the GPT-4o model, averaged over all test sets. Both example-based methods scale linearly with dataset size: Examples at 236 tokens per example and Summary-Ex at 126 tokens per example, reaching 25K and 12.5K tokens, respectively, at 100 examples. In contrast, our methods show minimal correlation with example count: **Description-based approaches remain under 200 tokens while Criteria-based methods stabilize under 600 tokens regardless of sample size, representing a reduction of up to 99% in token usage.**

This token efficiency translates directly to faster test-time processing (Figure 14). While the Examples approach requires over 300 seconds at 100 examples, the criteria and description methods consistently deliver 70% and 57% reductions, respectively. Summary-Ex shows moderate efficiency gains, maintaining reasonable performance up to 50 examples before scaling substantially. These results reveal that our proposed methods significantly improve efficiency while preserving performance.

7 Conclusions

This work enhances the efficacy of ICL by introducing novel knowledge extraction methods that distill examples into precise task instructions. We contribute a new benchmark spanning five essential B2B concepts, offering a valuable resource for advancing classification in professional contexts. Our dataset reveals the inherent complexity of B2B language understanding, where traditional and complex few-shot methods degrade sharply as context length increases, highlighting the need for more sophisticated approaches.

Through extensive experiments across varied LLMs and few-shot configurations, our methods demonstrate superior performance, time efficiency, and cost-effectiveness compared to the standard ICL and various token compression techniques. The proposed framework exhibits strong adaptability across different business concepts without requiring domain-specific customization. Crucially, our methods generate interpretable artifacts that enable seamless human-in-the-loop collaboration. This interpretability, combined with our flexible architecture that allows knowledge distillation from powerful to efficient models, opens new pathways for transparent and scalable NLP applications.

Limitations

The scope and implications of our research are bound by several limitations.

Although our experimental evaluation is grounded in a carefully curated resource of 50 B2B sales calls, each lasting 30 to 90 minutes and annotated by three independent professional annotators under expert supervision, broader validation across additional B2B corpora remains necessary to fully establish the robustness of our findings. The current scope reflects the substantial investment in high-quality human annotation, as we prioritize expert-annotated labels over automated alternatives.

Furthermore, our knowledge extraction framework relies on the reasoning capabilities of the underlying LLM to distill accurate criteria from verbose examples. Consequently, our method is susceptible to error propagation if the extraction model hallucinates incorrect rules or misses subtle semantic nuances. While our results show performance gains, the quality of the generated distilled artifacts is upper-bounded by the capability of the model used for distillation.

Finally, our work focuses on binary classification tasks. While our methods can easily be extended to multi-class classification, the extracted knowledge representation would increase linearly with the number of classes, potentially creating efficiency concerns. Furthermore, as the number of classes increases, some may be underrepresented in the sampled examples, limiting the quality of the generated criteria and descriptions. This may compromise generalization capabilities and overall performance in multi-class scenarios.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, and 1 others. 2024. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Model card.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. In *First Workshop on Long-Context Foundation Models@ ICML 2024*.

Marko Bohanec, Marko Robnik-Šikonja, and Mirjana Kljajić Borštnar. 2017. Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. *Organizacija*, 50(3):217–233.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. 2003. Expertise identification using email communications. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, page 528–531, New York, NY, USA. Association for Computing Machinery.

Joyce Chai, Jimmy Lin, Wlodek Zadrozny, Yiming Ye, Margo Stys-Budzikowska, Veronika Horvath, Nanda Kambhatla, and Catherine Wolf. 2001. The role of a natural language conversational interface in online sales: a case study. *International Journal of Speech Technology*, 4:285–295.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Andrew Kristoffer Dean, Nick Ellis, and Victoria K. Wells and. 2017. Science 'fact' and science 'fiction'? Homophilous communication in high-technology B2B selling. *Journal of Marketing Management*, 33(9-10):764–788. Publisher: Routledge _eprint: https://doi.org/10.1080/0267257X.2017.1324895.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods*

in Natural Language Processing, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Rajdeep Grewal, Gary L Lilien, J Andrew Petersen, and Stefan Wuyts. 2022. Business-to-business marketing: Looking back, looking forward. In *Handbook of business-to-business marketing*, pages 2–11. Edward Elgar Publishing.

Barbara Grosz, D Scott, H Kamp, P Cohen, and E Giachin. 1995. Discourse and dialogue.

Deepa Gupta, Anil Nair, Deepa Iyer, and Anil Reddy. 2022. Leveraging machine learning and natural language processing for enhanced b2b marketing intelligence. *Innovative AI Research Journal*, 11(10).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. 2025. ETHIC: Evaluating large language models on long-context tasks with high information coverage. 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 5497–5512, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Itay Lieder, Meirav Segal, Eran Avidan, Asaf Cohen, and Tom Hope. 2019. Learning a Faceted Customer Segmentation for Discovering new Business Opportunities at Intel . In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6136–6138, Los Alamitos, CA, USA. IEEE Computer Society.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand. Association for Computational Linguistics.

Amit Patel, Neha Patel, Rohit Reddy, and Priya Bose. 2022. Enhancing sales efficiency through ai: Leveraging natural language processing and reinforcement learning for automated sales tools. *International Journal of AI ML Innovations*, 11(8).

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*.

D. Rohaan, E. Topan, and C.G.M. Groothuis-Oudshoorn. 2022. Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. *Expert Systems with Applications*, 188:115925.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, and 23 others. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022 - Tenth International Conference on Learning Representations*, Online, Unknown Region.

Marco Spruit, Marcin Kais, and Vincent Menger. 2021. Automated Business Goal Extraction from E-mail Repositories to Bootstrap Business Understanding. *Future Internet*, 13(10):243.

Gianmario Voria, Francesco Casillo, Carmine Gravino, Gemma Catolino, and Fabio Palomba. 2024. Recover: Toward the automatic requirements generation from stakeholders' conversations. *arXiv preprint arXiv:2411.19552*.

Zhong Wu, Qiping She, and Chuan Zhou. 2024. Intelligent customer service system optimization based on artificial intelligence. *J. Organ. End User Comput.*, 36(1):1–27.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. CompAct: Compressing retrieved documents actively for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21424–21439, Miami, Florida, USA. Association for Computational Linguistics.

Eelaaf Zahid, Yuya Jeremy Ong, Aly Megahed, and Taiga Nakamura. 2021. Predicting Loss Risks for B2B Tendering Processes . In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2076–2083, Los Alamitos, CA, USA. IEEE Computer Society.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

The supplementary material is organized as follows: Appendices A, B, and C detail data curation, licensing, ethical governance, anonymization procedures, and dataset statistics. Appendices D and E describe the classification structure and knowledge extraction methodology. Appendix F presents cross-model knowledge distillation experiments, demonstrating how our methods can transfer knowledge from large-scale models to smaller ones. Appendices G, H, I, and J provide thematic analyses: comparing classification criteria versus descriptions, evaluating generated descriptions through the lens of abstraction versus example coverage, qualitatively assessing generated criteria, and illustrating human-in-the-loop modifications. Finally, Appendix K contains average processing times and Appendix L presents the comprehensive experimental performance grids.

A Data Curation and Usage Guidelines

A.1 Data Licensing and Intended Use

The dataset is made available for non-commercial research purposes in NLP. The complete license agreement will be provided as part of the data distribution package on the project website. All existing artifacts were used in accordance with their original intended purposes and licensing terms. Derivatives of this dataset must remain within research contexts to maintain compatibility with the original access conditions.

A.2 Ethical Approval and Governance

The data collection protocol underwent internal review by the appropriate legal and scientific governance bodies. This process ensured compliance with organizational ethical guidelines and data protection standards.

A.3 Annotator Profile and Recruitment

The annotation task was completed by female native speakers of American English. All annotators

were recruited through established professional networks and compensated at rates consistent with industry standards for linguistic annotation work in the United States.

B Data Anonymization

To enable public release, we applied a rigorous anonymization process.[5]

A trained annotator identified potentially sensitive information across all snippets, following guidelines covering personal names, organizations, products, locations, contact information, and numeric identifiers. We supplemented manual review with automated heuristics to detect capitalized terms and numeric patterns, and employed Claude Sonnet 3.7 as an additional safety check to identify any overlooked entities.

All identified entities were replaced with fictional alternatives while preserving semantic coherence. For numerical data, we substituted original values with randomized numbers from similar ranges, maintaining local consistency where needed. Professional roles, titles, and percentage values remained unchanged to preserve contextual meaning.[6]

As a final step, we simplified each snippet through controlled sentence-level rewriting using Claude Sonnet 3.7. We processed entire snippets at once to maintain conversational context, while instructing the model to rewrite individual sentences. We designed a specific prompt directing the model to preserve the semantic content while altering the syntactic structure and word choice. This ensured the text retained its original meaning and conversational flow while no longer resembling the original style or phrasing.

Figure 4 presents the prompt template used for the transcript simplification process. The input for this prompt is the text snippet that needs to be simplified. The text snippet is divided into enumerated sentences. The prompt facilitates controlled sentence-level rewriting while preserving the semantic content of conversations. When applying this prompt, the LLM rewrites each line individually, maintaining the conversation's structure and essential meaning.

[5]Experiments confirmed no performance degradation between original and anonymized data.

[6]Our repository provides mappings from entity types to the set of fictional replacements used.

```
<instructions>
Please do the following based on the text
in <snippet></snippet> tags:
- Rewrite the text in each row separately
in a simplified, short, and clear way,
maintaining the original phrasing as
closely as possible.
- Rewrite the text in each row separately,
ensuring that each rewritten row
corresponds exactly to the same row in the
original snippet. DO NOT skip any rows, and
do not combine content from multiple rows
into one.
- You should retain all the important
details of the transcript.
</instructions>

<snippet>
{TEXT_SNIPPET}
</snippet>

<format>
x | SPEAKER_AFFILIATION Rewritten text for
row number x
y | SPEAKER_AFFILIATION Rewritten text for
row number y
...

<example>
**Original Transcript:**
1 | [PROSPECT_A] I have a, I think, a
meeting at 10 AM. So I need to... Wait a
minute.
2 | [PROSPECT_A] So.
3 | [PROSPECT_A] Let me see if, let me
check with the team. I will need to confirm
it with them.
4 | [SELLER_A] OK, I mean, the, the, the
presentation is due by Friday.

**Rewritten Transcript:**
1 | [PROSPECT_A] I have a 10 AM meeting.
2 | [PROSPECT_A] So.
3 | [PROSPECT_A] let me check and confirm
with the team.
4 | [SELLER_A] The presentation is due by
Friday.
</example>
</format>
```

Figure 4: Prompt template used for transcript simplification.

C Dataset Statistics and Example Snippets

Table 3 summarizes key statistics for the `Call Playbook` dataset, including the number of positive and negative examples, total sample count, number of unique calls, average number of words per sample, and the proportion of dialogues attributed to the prospect. The dataset maintains a balanced or nearly balanced distribution across most categories and contains detailed conversational examples of

business interactions.

Table 4 presents representative positive snippets from each of the five concepts, highlighting distinct business aspects reflected in the dataset. The bolded portions highlight the most relevant spans for each category. Colors distinguish between prospect (orange) and seller (blue) utterances. These examples, processed and anonymized according to our procedures, reflect the rich diversity and complexity of B2B dialogues.

D Classification Methodology

This section details our approach to classifying B2B conversation snippets, including the classification prompt structure, user intents, and knowledge section formats used in our experiments.

D.1 Classification Prompt Structure

Our classification system implements a structured prompt template consisting of three main components: (1) a classification objective incorporating the user-provided intent, (2) a knowledge section derived from labeled examples, and (3) the desired structured output format.

Figure 5 presents the base classification prompt template, which remains consistent across all variations. The prompt instructs the model to analyze a text snippet and determine whether it contains evidence of a specific B2B concept (e.g., "the prospect discusses their business goals"). The user intents used in our experiments are detailed in Section D.2. Variation across prompts arises from the knowledge section, which includes either examples, summarized examples, criteria, or descriptions, as discussed in Section D.3. The prompt concludes with the text snippet to be classified and specifies the expected response format.

D.2 User Intents

We conducted our experiments using five distinct classification objectives, each based on a user-provided intent aligned with one of the five B2B conversational concepts:

- **Business Goals:** "the prospect discusses their business goals in the context of the purchasing process".
- **Decision Criteria:** "the prospect discusses their decision criteria in the context of the purchasing process".

```
<instructions>
Analyze the following text snippet and
identify whether {USER_INTENT}.

Provide a detailed reasoning for your
decision (chain of thoughts) before
delivering the final classification.

Label the snippet as either Positive (if
{USER_INTENT}) or Negative (if the snippet
does not relate or contain relevant
information).

{KNOWLEDGE_SECTION}

Base your classification on the objective
and the {KNOWLEDGE_TYPE} provided above.
</instructions>

<snippet>
{TEXT_SNIPPET}
</snippet>

<format>
<rationale> [Your reasoning] </rationale>
<label> [Positive or Negative] </label>
</format>
```

Figure 5: Base classification prompt template.

- **Decision Makers:** "the prospect mentions the decision makers involved in the purchasing process".
- **Decision Making Process:** "the prospect discusses their decision-making process regarding the purchase".
- **Pain Points:** "the prospect discusses their pain points in the context of the purchasing process".

These intents were inserted into the classification prompt template at the `{USER_INTENT}` placeholder shown in Figure 5.

The intents are intentionally high-level, reflecting the typical level of specificity provided by users in real-world applications. This level of abstraction in such intents is common, as users may not always possess the domain expertise or technical vocabulary to formulate precise classification parameters. This limitation further motivates our approach of augmenting user-provided intents with knowledge derived from labeled examples.

D.3 Knowledge Section Formats

As described in Section D.1, we implemented three distinct formats for the `{KNOWLEDGE_SECTION}`

	Train						Test					
Dataset	**Pos**	**Neg**	**Total**	**Calls**	**Avg Words**	**Prospect %**	**Pos**	**Neg**	**Total**	**Calls**	**Avg Words**	**Prospect %**
Business Goals	100	100	200	25	288	48	100	100	200	25	279	51
Decision Criteria	86	114	200	25	277	50	94	106	200	25	274	47
Decision Makers	32	168	200	25	259	48	35	165	200	25	274	43
Decision Making Process	61	139	200	20	271	48	68	132	200	21	262	42
Pain Points	100	100	200	25	275	45	100	100	200	25	296	49

Table 3: `Call Playbook` statistics. Pos: positive examples; Neg: negative examples; Total: sample count; Calls: number of unique calls; Avg Words: average words per sample; Prospect %: percentage of prospect-side dialogue per sample.

Dataset	**Snippet Example**
Business Goals	[PROSPECT_A] Yeah. I'm head of advertising and analytics. Most of our sales are retail. SchistHorizon Networks is our eCommerce star. **We're trying to increase sales through Omega Operations channel, using Triton Trades for attention and conversions.** GoldenLeaf Enterprises seems to fit our audience: mostly female, 90-30 split, slightly older millennials. So... yeah, that's... [SELLER_A] Okay. Your site looks nice. What inspired this company? [PROSPECT_A] Yeah. We make bottled water and coffee, exploring other options too. Our focus is sustainability. Our source is in Westvale. [SELLER_A] Okay. [PROSPECT_A] Our motto is "premium by nature". We don't use chemicals for alkaline water. It's naturally alkaline due to volcanic filtration. That's the core of everything we do. [SELLER_A] Hey, Jean. [PROSPECT_B] Hi there. Sorry I'm late. I was held up on another call but I'm excited to learn about GoldenLeaf Enterprises.
Decision Criteria	[SELLER_A] Yeah, yeah. Based on historical preferences, we want to discuss a ChalkForest Solutions option for you to position ourselves best. This allows them to do it their preferred way. If they change their mind, that's okay. We want you to be prepared for that. I updated Max before our call about our discussions, Val, including your ongoing proof of concept. He's now up to date on our progress. [PROSPECT_A] Your competitors provided four quotes: monthly and three-year contract options. If you can provide a second quote, that's great because you'll be competing on both options, and I can present choices to management. They can choose either option, and this company can offer both. Now it's about what we want. **We can discuss which company we prefer, focusing on features, support, and functionality.** Do we like the phones? Or not like the phones, things like that? And get into the detailed aspects? [SELLER_B] Well, I think that from a financial perspective.
Decision Makers	[PROSPECT_A] That's awesome. Very cool. We'd love to see what's involved. Are there any fees for us to use these services? Or is it just? [SELLER_A] No, it's completely complementary to you. LimitlessLogic was created because companies saw employees using it and spending money. The idea was to provide a business experience similar to LimitlessLogic, encouraging personal use. [PROSPECT_A] Yes. [SELLER_A] Yes. That's how this all came about. We used to charge a 10 percent fee before COVID, but we removed it. It's now completely complementary to your organization. We can set it up on LimitlessLogic and VortexVault, customizing programs for your company. [PROSPECT_A] Sounds good. Excellent. **We'd love to see how to do this, probably involving our people team.** Okay? They would be the ones to roll it out to the company. If you could send that information to me, that would be fantastic. [SELLER_A] I'll send a follow-up email with the information we discussed. Do you have your calendar available to schedule a demo? Are you available next week? What works best for you?
Decision Making Process	[SELLER_A] No problem. Swapping is not an issue. If someone leaves the company, we can swap them immediately and easily. I do it on our end. If an analyst leaves next month, we can add someone else in April without any problem. Kim and I can easily add them and transfer the saved content. If you move teams, we can transfer all your saved content to your replacement. We do this with all clients. We understand people change roles or leave companies. This is not limited. If someone leaves tomorrow and their replacement leaves in two weeks, No problem. We can swap them. We can't allow sharing because it's hard for us to approve. [SPEAKER_A] Yes. I understand. **I need to give this feedback to my partners because the three-process was ineffective**. We like the platform and want to continue as explained. But for a user with a laptop, we need to make a console. We can do it, but we're reducing to four laptops. Now, the four-process is still more ineffective. We need to decide if that works for us.
Pain Points	[SELLER_A] What specifically? OIN is failing and it's an issue now. Is there a reason for the urgency? Is it within the next month? [PROSPECT_A] A workshop over the next quarter. I want to solve the password management problem soon. Yeah, I... [SELLER_A] What happens if you don't? Is it just a personal goal? But nothing else? [PROSPECT_A] If solved, clients get a better user experience. **I have an inefficient account management team. Half my engineers are fixing password problems constantly.** For the business case, It's about efficiency, and user experience. So people can log in and take their training. Currently, people struggle to log in for training. [SELLER_A] How many people use this monthly? How many should log in versus how many actually do?

Table 4: Representative positive examples from the `Call Playbook` Dataset. Relevant spans are bolded.

placeholder of the classification prompt: the Examples format (also used by Summary-Ex) shown in Figure 6, the Criteria format shown in Figure 7, and the Description format shown in Figure 8.

For all formats, positive examples or criteria precede negative ones, as early experiments confirmed that this ordering produced superior performance.

```
Below is a list of positive examples that
would indicate that the objective is
present in the text snippet:
<positive_examples>
{POSITIVE_EXAMPLES}
</positive_examples>

Below is a list of negative examples that
would not indicate that the objective is
present in the text snippet:
<negative_examples>
{NEGATIVE_EXAMPLES}
</negative_examples>
```

Figure 6: Example-based knowledge guidance format that uses a direct few-shot approach with labeled examples from the dataset.

```
Below is a list of positive criteria that
would indicate that the objective is
present in the text snippet:
<positive_criteria>
{POSITIVE_CRITERIA}
</positive_criteria>

Below is a list of negative criteria that
would not indicate that the objective is
present in the text snippet:
<negative_criteria>
{NEGATIVE_CRITERIA}
</negative_criteria>
```

Figure 7: Criteria-based knowledge guidance format where an LLM distills labeled examples into explicit classification criteria.

```
Below is a detailed description of the
classification task:
<description>
{DESCRIPTION}
</description>
```

Figure 8: Description-based knowledge guidance format where an LLM generates a comprehensive explanation of the classification task based on labeled examples.

E Knowledge Extraction Process

Our approach transforms raw labeled examples into condensed, structured knowledge representations using LLMs. This section introduces the three types of knowledge representations used in our prompts: summaries, criteria, and descriptions. For each type, we describe the corresponding prompt design and the process used to generate it from labeled examples. Each representation was then inserted into one of the knowledge section formats described in Appendix D.3.

E.1 Summary Generation

To generate summarized examples for the Summary-Ex method, we applied a text summarization process using Claude Sonnet 3.7 that compresses the original labeled snippets while preserving their essential meaning, speaker structure, and conversational flow. Figure 9 shows the prompt template used to generate the summaries.

This approach addresses potential issues with lengthy examples by creating condensed versions that maintain core conversational patterns and sales concepts. The goal is to reduce prompt length while maintaining key business content and discourse patterns. This method aims to reduce overall prompt length while preserving essential information, potentially improving the model's focus on relevant content patterns.

The resulting summaries replace the full-length examples in the Examples format (Figure 6).

E.2 Criteria Generation

To generate classification criteria, we employed two variants. The first variant, Criteria-Ex, derives criteria directly from examples, while the second variant, Criteria-De, derives criteria from an existing task description (generated by Description-Ex). For this purpose, we employed a prompt template that guides the relevant model to generate two lists: positive criteria that indicate the presence of the target concept and negative criteria that indicate its absence. Figure 10 shows the prompt template used for this purpose.

In both variants, the prompt emphasizes the need for general, clear, and concise criteria that can be applied to any text snippet. In the Criteria-Ex variant, the model is further instructed to base each criterion on patterns observed in at least two of the provided examples. For zero-shot prompting, we omit the `{KNOWLEDGE_SECTION}` and generate the

```
<instructions>
Analyze the following B2B call text snippet
and create a simplified, concise version of
the original text.
Preserve the original format and maintain
all speaker affiliations exactly as they
appear in the source.
Focus on removing redundant information and
filler words while keeping all discussed
content, focusing on main business topics.
Keep the essential structure and flow of
the conversation intact.

Create a condensed version that captures
what was discussed without changing the
text format or speaker affiliations.
The summary should include 3-5 sentences at
most.
</instructions>

<snippet>
{TEXT_SNIPPET}
</snippet>

<format>
Your answer must be in the following format:
<summary>
[Simplified version of the original text]
</summary>
</format>
```

Figure 9: Summary generation prompt template.

criteria based solely on the user intent.

Both variants follow the same knowledge section formats outlined in Appendix D.3, as illustrated in the prompt template shown in Figure 10, adapting them to the specific requirements of criteria generation by incorporating the appropriate content into the {KNOWLEDGE_SECTION} placeholder, where the {KNOWLEDGE_TYPE} can be either "examples" or "description" depending on the specific variant being employed.

The resulting criteria are then inserted into the knowledge section embedded within the classification prompt, using the criteria variant (Figure 7).

E.3 Description Generation

To generate task descriptions, we similarly employed two variants. The first variant, Description-Ex, derives a description directly from examples, while the second variant, Description-Cr, derives a description from existing criteria (generated by Criteria-Ex). To this end, we employed a prompt template that guides the relevant model to produce a detailed description of the classification task. Figure 11 shows the prompt template used for description generation.

```
<instructions>
You are tasked with annotating text
snippets.

The end-goal task is to analyze text
snippets and determine whether
{USER_INTENT}.

Your task is to generate two lists:
A list of positive criteria: Conditions that
indicate the presence of the objective.

A list of negative criteria: Conditions
that indicate the absence of the objective.

{KNOWLEDGE_SECTION}

Base your criteria on the objective and the
{KNOWLEDGE_TYPE} provided above.
The criteria should be as general as
possible and should be applicable to any
text snippet.
The criteria should be clear and concise.
Each list of criteria should include at
least five criteria and no more than ten
criteria.
Each criterion should be self-explanatory
and not require an example.
[For Criteria-Ex: Each criterion should be
based on at least two of the examples
provided above.]
</instructions>

<format>
Your answer must be in the following format:
<criteria>
<positive>
Criterion 1: [Criterion 1]
Criterion 2: [Criterion 2]
...
</positive>
<negative>
Criterion 1: [Criterion 1]
Criterion 2: [Criterion 2]
...
</negative>
</criteria>

[Example format omitted for brevity]
</format>
```

Figure 10: Criteria generation prompt template.

In both variants, the prompt instructs the model to generate a description of the classification task that facilitates effective recognition of the target concept in conversational text. The instructions emphasize the importance of generating descriptions that strike a balance between comprehensiveness and brevity, ensuring they can be applied consistently across diverse text samples. For zero-shot prompting, we omit the {KNOWLEDGE_SECTION} and generate the description based solely on the

user intent.

Both variants utilize the knowledge section formats outlined in Section D.3, as illustrated in the prompt template shown in Figure 11, adapting them to the specific requirements of description generation by incorporating the appropriate content into the {KNOWLEDGE_SECTION} placeholder, where the {KNOWLEDGE_TYPE} can be either "examples" or "criteria" depending on the specific variant being employed.

The resulting description is then embedded within the classification prompt, using the description variant (Figure 8).

```
<instructions>
You are tasked with annotating text
snippets.

The end-goal task is to analyze text
snippets and determine whether
{USER_INTENT}.

Your task is to generate a detailed
description of the classification task
allowing for the identification of the
objective in text snippets.

[KNOWLEDGE SECTION]

Base your description on the objective and
the {KNOWLEDGE_TYPE} provided above.
The description should be as general as
possible and should be applicable to any
text snippet.
The description should be clear and concise.
</instructions>

<format>
Your answer must be in the following format:
<description>
[Your description]
</description>

[Example format omitted for brevity]
</format>
```

Figure 11: Description generation prompt template.

F Cross-Model Knowledge Distillation

Unlike standard few-shot learning, our knowledge extraction methods allow us to leverage the generation capabilities of larger, more capable models to enhance the classification performance of smaller models. This approach requires only a single generation step, with minimal time or cost overhead.

We generate criteria and descriptions with the larger models (Sonnet 3.7, Mistral Large) and inject them into the in-context prompts of the smaller models (Haiku 3, Mistral Small) for classification.

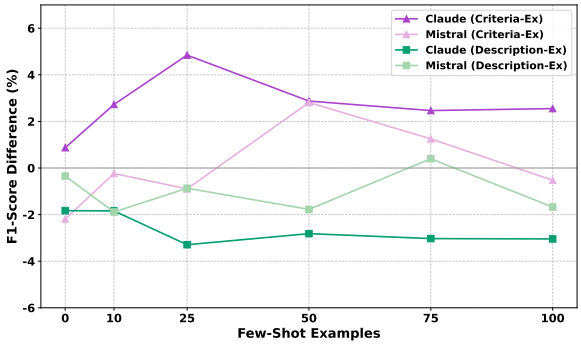

Figure 12: Macro-average F1 difference when applying criteria and descriptions generated by larger versus smaller models. Classification was performed by smaller models in all cases. Positive values indicate better performance with larger model-generated content.

Figure 12 illustrates the macro-average F1 difference, averaged over all five concepts, between content generated by the larger models and the original content produced by the smaller models. The results reveal a clear pattern: smaller models consistently improve when using structured criteria from larger models, but perform worse when incorporating descriptions generated by larger models. Mistral models demonstrate lower variance compared to Claude models, though both model families follow the same overall trend. This suggests that **structured criteria transfer more effectively between models than verbose descriptions**. The rule-based, concise nature of criteria appears more universal than verbose text, which can be highly specific to the source model's generation style, enabling smaller models to better leverage the distilled conceptual structure.

G Comparative Analysis of the Criteria and Description Methods

This appendix presents a comparative analysis of the four criteria and description methods. Tables 5 and 6 show the texts generated by Claude Sonnet 3.7 using 25 few-shot examples from our first iteration, applied to the Business Goals concept.

To illustrate how these task instructions function in practice, we analyze their application to two positive and two negative representative examples:

- Positive 1: A conversation in which the prospect articulates how they view their customer service channel as a revenue opportunity, referencing specific metrics such as

call handling times and outlining their business goals for transforming it into a more marketing-oriented function.

- Positive 2: A conversation in which the prospect describes content management challenges, specifically their need to maintain consistency across product documentation and how centralized updates would improve operational efficiency.

- Negative 1: A conversation in which the participants focus entirely on small talk about an earthquake and personal topics, with no articulation of business objectives.

- Negative 2: A conversation in which the participants focus solely on contract terms and technical settings between sellers, without any expression of the prospect's business goals or strategic needs.

The side-by-side format in both tables separates the instructional content (criteria definitions or descriptive guidance) from the illustrative examples, making it easier to see how these examples influenced the generation of the criteria and descriptions, and showing the different approaches to identifying business objectives in sales conversations.

Our analysis reveals several key distinctions and similarities between the four instructional approaches:

Structural Differences: The criteria-based approaches (Criteria-Ex and Criteria-De) provide discrete, categorical guidelines that annotators can apply systematically. In contrast, the narrative descriptions (Description-Ex and Description-Cr) offer more contextual guidance and flow more naturally, potentially making them more accessible to non-expert annotators.

Emphasis Variations: Criteria-Ex emphasizes how the prospect relates a product or service to their business context, while Criteria-De focuses more on the nature of the business objectives themselves. Description-Ex highlights conversational indicators of business goals, whereas Description-Cr emphasizes measurable outcomes and operational insights.

Complementary Coverage: All four approaches effectively identify the positive examples as containing business objectives, but through different analytical lenses. For instance, Positive 1 is recognized through explicit goal statements in Criteria-Ex but through measurable metrics in Criteria-De.

Negative Case Identification: Each approach successfully flags the negative examples as lacking business objective articulation, but with different emphasis: Negative 1 is identified primarily through its small-talk nature, while Negative 2 is flagged for its focus on administrative details without business context.

Granularity vs. Holistic Assessment: The criteria-based approaches offer more granular analytical points, potentially supporting more consistent annotation across different raters. The narrative descriptions provide a more holistic framework that may better capture contextual nuances in articulating business objectives.

This analysis demonstrates how our task instructions guide annotators through different analytical pathways. While criteria-based methods enable systematic decomposition of conversational elements, narrative descriptions encourage comprehensive evaluation of speaker intent. These findings inform the design of robust annotation protocols that balance analytical precision with contextual sensitivity in conversational analysis.

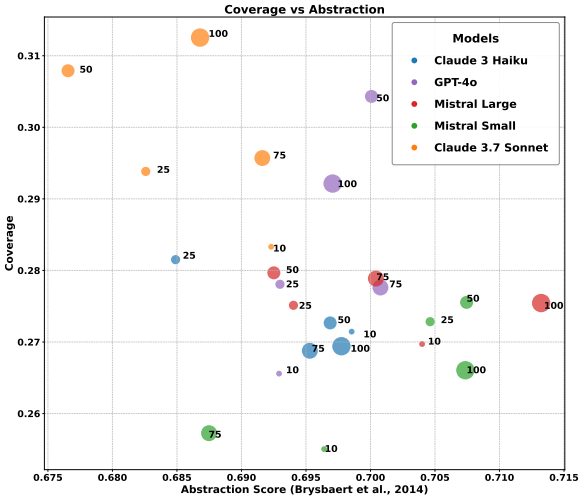

Figure 13: Trade-off between abstraction and coverage in task descriptions. Point size indicates the number of examples used. Higher abstraction scores reflect more abstract descriptions, while higher coverage indicates a better representation of examples.

H Abstraction vs. Coverage in LLM-Generated Descriptions

In Figure 13, we examine how LLMs balance coverage and abstraction through the descriptions generated by Description-Ex. Coverage is measured

Criteria-Ex: Positive		
ID	**Criteria**	**Illustrative Quotes**
Crit. 1	The prospect explicitly states their business objectives or goals that they hope to achieve through the purchase or implementation of the product/service	**"We want that channel to be more marketing-minded and product-focused rather than service-focused."**
Crit. 2	The prospect discusses how the product/service would integrate with or improve their existing business processes, workflows, or operations	**"We wonder if it's possible to maintain all this data in a single place... so updating one place automatically updates all documents."**
Crit. 3	The prospect explains specific business challenges or pain points they are trying to solve through the purchasing decision	**"When we change one document, it must be changed in all documents. That's our current challenge."**
Crit. 6	The prospect shares information about their business model, operational structure, or customer relationships in the context of how the purchase would impact these areas	**"Most calls are about order tracking, which we can address with self-service... When a rep is on a product inquiry call, about six other customers are waiting for service."**
Criteria-Ex: Negative		
Crit. 2	The prospect discusses pricing, contracts, or payment terms without relating them to broader business goals or outcomes	**"They're also in a two-year contract. If they add seats, they'd pay for six years, right?"**
Crit. 5	The conversation consists primarily of small talk, introductions, or unrelated topics that don't touch on business objectives	**"Everyone's talking about the Ironwood earthquake over the last 25 minutes here on the Corswick. It was in Ivorycliff."**
Crit. 7	The conversation is dominated by the seller explaining their offering without the prospect articulating how it connects to their business objectives	**"I'll ask if they want 20 seats or remind them account sharing isn't allowed. We can let them out of the contract. If they stay, they need more users."**
Criteria-De: Positive		
Crit. 2	The prospect describes specific problems or pain points in their current business operations that they are looking to solve through the purchasing decision	**"Between reports, about 24 percent of the data is common... When we change one document, it must be changed in all documents."**
Crit. 3	The prospect articulates measurable targets, metrics, or key performance indicators they aim to improve through the acquisition of a product or service	**"A customer inquiring about a product takes 6 to 20 minutes to close. A customer service call takes about two minutes... When a rep is on a product inquiry call, about six other customers are waiting for service."**
Crit. 6	The prospect outlines specific operational efficiencies, cost savings, or productivity improvements they expect to gain from the purchase	**"One advantage we see is building blocks, as you mentioned. Updating one block should update all reports."**
Crit. 7	The prospect connects features or capabilities of the product/service directly to their business needs or organizational priorities	**"We see an opportunity because we have a wide range of products that require education about fit. For our women's assortment, we have about seven different fits, multiplied by 25 fabric types, colors, and washes. It's complex."**
Criteria-De: Negative		
Crit. 2	The conversation consists primarily of the seller explaining potential benefits without the prospect articulating their own business goals or needs	**"I'll ask if they want 20 seats or remind them account sharing isn't allowed. We can let them out of the contract. If they stay, they need more users."**
Crit. 4	The text contains only small talk, pleasantries, or relationship-building conversation unrelated to business goals or purchasing decisions	**"You hear me? Yeah, hi, Kenzie. How are you?... I'm in Quorvath... My family is on the Corswick... My sister is flying to Torrengard tonight. She plays lacrosse."**
Crit. 5	The prospect discusses only pricing, contract terms, or payment options without relating these to their business objectives or expected outcomes	**"They complained, so we set them to two for nine users. They're on two now."**

Table 5: Comparative Analysis of Criteria-Based Instructions: Highlighted quotes illustrate how positive examples (blue/teal) inform the positive criteria and negative examples (red/brown) inform the negative criteria. The table demonstrates how structured criteria from both the Criteria-Ex and Criteria-De approaches capture different aspects of the Business Goals concept.

Description-Ex	
Description	**Illustrative Quotes**
The task is to determine whether any part of the text shows the prospect discussing their business goals, objectives, or desired outcomes in relation to a purchasing decision or process	**"I definitely see it as an opportunity... I see it as a sales channel."** **"We have multiple products and need to prepare reports for each one... That's our current challenge."**
This includes when prospects explain what they want to achieve with a product/service. Prospects may describe their organization's strategic aims or outline operational needs	**"We want that channel to be more marketing-minded and product-focused rather than service-focused."** **"So updating one place automatically updates all documents, like for a plant manufacturing 6 products."**
Look for instances where prospects articulate their vision, priorities, requirements, or expected benefits from implementing a solution	**"We see an opportunity because we have a wide range of products that require education about fit."** **"One advantage we see is building blocks, as you mentioned. Updating one block should update all reports."**
This may involve discussions about improving processes, solving problems, or enhancing efficiency	**"When we change one document, it must be changed in all documents. That's our current challenge... We wonder if it's possible to maintain all this data in a single place."** **"Most calls are about order tracking, which we can address with self-service and more resources."**
Exclude general small talk or discussions where only the seller is talking without the prospect articulating their own business objectives	**"Everyone's talking about the Ironwood earthquake over the last 25 minutes here on the Corswick."** **"I'll ask if they want 20 seats or remind them account sharing isn't allowed. We can let them out of the contract. If they stay, they need more users."**
Description-Cr	
The task is to determine whether any part of the text shows the prospect articulating their business goals, objectives, or desired outcomes in relation to a potential purchase or implementation	**"I definitely see it as an opportunity... I see it as a sales channel. We want that channel to be more marketing-minded and product-focused rather than service-focused."** **"We wonder if it's possible to maintain all this data in a single place... Updating one block should update all reports."**
Look for instances where the prospect connects the purchase decision to measurable business outcomes, ROI expectations, or organizational growth plans. The objective is present when prospects share insights about their business model or operational structure	**"A customer inquiring about a product takes 6 to 20 minutes to close. A customer service call takes about two minutes... When a rep is on a product inquiry call, about six other customers are waiting for service."** **"We have a wide range of products that require education about fit. For our women's assortment, we have about seven different fits, multiplied by 25 fabric types, colors, and washes. It's complex."**
Exclude conversations that focus solely on technical specifications, pricing details, or administrative aspects without connection to broader business goals. Also exclude instances when the conversation is dominated by the seller without the prospect articulating how the offering aligns with their business objectives, or contains only pleasantries	**"They're also in a two-year contract. If they add seats, they'd pay for six years, right?... They complained, so we set them to two for nine users."** **"I'll ask if they want 20 seats or remind them account sharing isn't allowed. We can let them out of the contract."** **"You hear me? Yeah, hi, Kenzie. How are you?... Everyone's talking about the Ironwood earthquake."**

Table 6: Comparative Analysis of Description-Based Instructions: Highlighted quotes illustrate how positive examples (blue/teal) inform the positive guidance and negative examples (red/brown) inform the exclusion guidance. The table demonstrates how contextual descriptions from both Description-Ex and Description-Cr approaches provide holistic frameworks for the Business Goals concept.

as the cosine similarity between model-generated descriptions and their source examples using *all-MiniLM-L6-v2* embeddings, while abstraction is quantified using the concreteness ratings from Brysbaert et al. (2014). Each point represents a model-dataset pairing with varying few-shot counts (indicated by point size), revealing how these properties interact across different dimensions.

The figure demonstrates a clear trade-off between these properties. Sonnet 3.7 achieves the highest coverage scores, particularly with larger example sets, but operates at lower abstraction levels. Conversely, both Mistral models demonstrate superior abstraction capabilities while exhibiting

reduced coverage of the original examples.

This pattern highlights an inherent tension: **descriptions that closely mirror their source examples tend to be more concrete, while more abstract descriptions capture fewer specific details from the training data.** GPT-4o presents a more balanced approach, achieving good abstraction while retaining reasonable coverage, especially when provided with moderate to large example sets.

Interestingly, while we expected more examples to increase abstraction and reduce coverage, only Mistral models follow this pattern. Sonnet 3.7 exhibits the opposite trend, and Haiku 3 shows no clear correlation, suggesting that these LLMs differ in their learning approaches.

I Qualitative Analysis of Generated Classification Criteria

To better understand the characteristics of the generated criteria for our B2B concepts, we conducted a systematic analysis of more than 3,000 criteria produced across all experimental conditions. We examined several key linguistic properties for each criterion:

- **Number of criteria**: The average number of positive and negative criteria generated per class, reflecting the model's ability to articulate classification rules.
- **Abstraction**: Computed using established concreteness ratings from the Brysbaert et al. (2014) lexical database, with higher scores indicating more abstract language.
- **Business indicators**: Identified using pattern matching against a comprehensive lexicon of business value terminology (e.g., ROI, customer retention, market share).
- **Implementation focus**: Detected through the presence of technical and procedural terminology related to system deployment and execution processes.
- **Solution orientation**: Assessed based on the presence of solution-focused verbs and outcome-oriented language.
- **Conditional logic**: Tracked through business-relevant conditional statements and logical constructions.

Table 7 presents the differences between criteria derived from examples versus those derived from descriptions, revealing distinct linguistic patterns. Our analysis reveals that Criteria-Ex exhibits higher rates of business value references (+5.7%) and implementation details (+10.0%) compared to Criteria-De. Both approaches yield identical high abstraction scores (0.69) and maintain comparable numbers of positive and negative criteria, with Criteria-De generating slightly more criteria per sample.

Property	Criteria-Ex	Criteria-De
Positive criteria	6.0	6.5
Negative criteria	6.0	6.5
Abstraction (0-1 scale)	0.69	0.69
Business value indicators	64.4%	58.7%
Implementation details	59.0%	49.0%
Solution-focused language	29.1%	24.8%
Conditional business logic	33.9%	38.1%
Quantifiable metrics	0.1%	0.0%

Table 7: Linguistic characteristics of generated classification criteria.

A notable finding is the near absence of quantifiable metrics across all criteria, suggesting that models prioritize qualitative over quantitative reasoning when establishing classification guidelines. Additionally, both approaches demonstrate relatively low usage of conditional logic, although Criteria-De employs conditional statements more frequently (38.1%) than Criteria-Ex (33.9%). This pattern indicates that the generated criteria tend to favor declarative statements over conditional or logical formulations.

J Human-in-the-Loop Modification Example

To illustrate the nature and impact of human modifications to our generated criteria (See Section 6.3), we present a detailed comparison of the original model-generated criteria of Sonnet 3.7 for the Pain Points classification task and the modifications made by two annotators with contrasting performance outcomes.

Table 8 shows selected criteria where both annotators made modifications, comparing the original Criteria-Ex output with the human-revised versions. We focus on criteria where substantive changes were made by both annotators to highlight the different modification strategies employed. The original model achieved 73.2% F1 on this task.

The table reveals two distinct modification approaches with contrasting outcomes. Annotator 1's modifications achieved 83.0% F1 (+9.8% improvement) through strategic simplification (remov-

Criterion	Claude Sonnet 3.7	Annotator 1 (+9.8% F1)	Annotator 2 (-1.3% F1)
Positive 1	The prospect explicitly describes challenges, frustrations, or difficulties they are experiencing with their current purchasing process, systems, or vendors.	The prospect explicitly describes challenges, frustrations, or difficulties they are experiencing with their current **solutions**, systems, or vendors.	The **potential customer openly discusses struggles**, frustrations, or issues they are encountering with their current purchasing process, systems, or **suppliers**.
Positive 2	The prospect mentions inefficiencies, wasted time, or resource constraints that are directly impacting their ability to make purchasing decisions or implement solutions.	The prospect mentions inefficiencies or **wasted time**.	The **potential customer** points out inefficiencies, wasted time, or resource limitations that are **directly affecting their capability to make purchase decisions or implement solutions**.
Negative 1	The prospect discusses general business information, company background, or role descriptions without mentioning any challenges or difficulties related to purchasing processes.	The prospect discusses general business information, company background, or role descriptions without mentioning any **challenges or difficulties**.	The **potential customer** discusses general business **data**, company **history**, or role descriptions without mentioning any **struggles or issues related to purchasing processes**.
Negative 6	The conversation focuses primarily on logistics, scheduling, or administrative details of the current meeting rather than business challenges.	*(No change)*	The conversation is primarily **centered** on logistics, scheduling, or administrative details of the current meeting rather than **business struggles**.

Table 8: Comparison of original model-generated criteria and human modifications for Pain Points classification. Bold text highlights key differences from the original version.

ing verbose phrases like "directly impacting their ability to make purchasing decisions"), semantic generalization (broadening "purchasing process" to "solutions"), and scope expansion (removing overly restrictive qualifiers). These changes enhanced the criteria's applicability and clarity.

Annotator 2's modifications resulted in 71.9% F1 (-1.3% degradation) and involved excessive paraphrasing without semantic improvement ("prospect" to "potential customer", "challenges" to "struggles"), increased linguistic complexity ("directly affecting their capability" vs "directly impacting their ability"), and inconsistent terminology introduction.

These findings demonstrate that effective modifications require strategic simplification, semantic broadening, and principled scope adjustments, while ineffective changes typically involve superficial rewording and unnecessary complexity. Importantly, our approach provides users with interpretable artifacts that can be meaningfully refined to improve classification performance through targeted human expertise.

K Average Processing Time

Figure 14 presents the average processing time of GPT-4o across all test sets, as referenced in Section 6.4. The processing times reported include the overhead of the knowledge extraction step (e.g., criteria, descriptions). It is important to note that for any given set of few-shot examples, our methods require only a single LLM call during the knowledge extraction phase, making this overhead minimal and required only once.

This visualization supports our findings regarding the computational efficiency advantages of our proposed methods compared to the traditional few-shot approach, especially as the number of examples increases. While Summary-Ex offers moderate improvements over the standard Examples method, the criteria and description methods scale more efficiently with increasing example counts, maintaining stable processing times regardless of the number of examples used.

L Comprehensive Experimental Results

L.1 Complete Performance Grid

Figure 15 illustrates the detailed evaluations of the macro-average F1 score, as discussed in Section 6.1. This visualization presents performance across five concepts (Business Goals, Decision Criteria, Decision Makers, Decision Making Process, and Pain Points) and five models (GPT-4o, Claude Sonnet 3.7, Claude Haiku 3, Mistral Large, and Mistral Small). The grid layout allows for direct

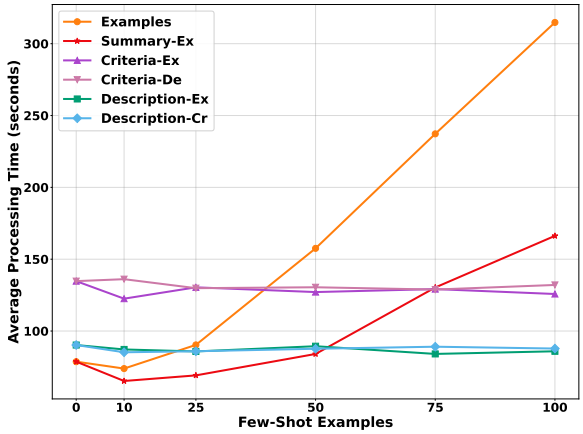

Figure 14: Average processing test time (in seconds) by method across different numbers of few-shot examples.

comparison of how each few-shot learning method performs as the number of examples increases from 0 to 100. Notably, the visualization confirms that the criteria and description methods generally maintain higher F1 scores than the traditional Examples approach and its summarized variant (Summary-Ex), both of which often exhibit declining performance with additional examples, a trend evident in all models except Sonnet 3.7.

L.2 Token Compression Baseline Comparison

Figure 16 presents the detailed per-dataset comparison with token compression methods, averaged across all five models. The results complement Table 1 in Section 6.2, showing consistent patterns across all concept extraction tasks. Both Criteria-Ex and Description-Ex demonstrate robust performance as examples increase, while LLMLingua-2 performs poorly around 50% F1 or below across all datasets as it compresses both task instructions and examples. SC shows competitive initial performance but exhibits consistent degradation, with steep declines observed in all tasks.

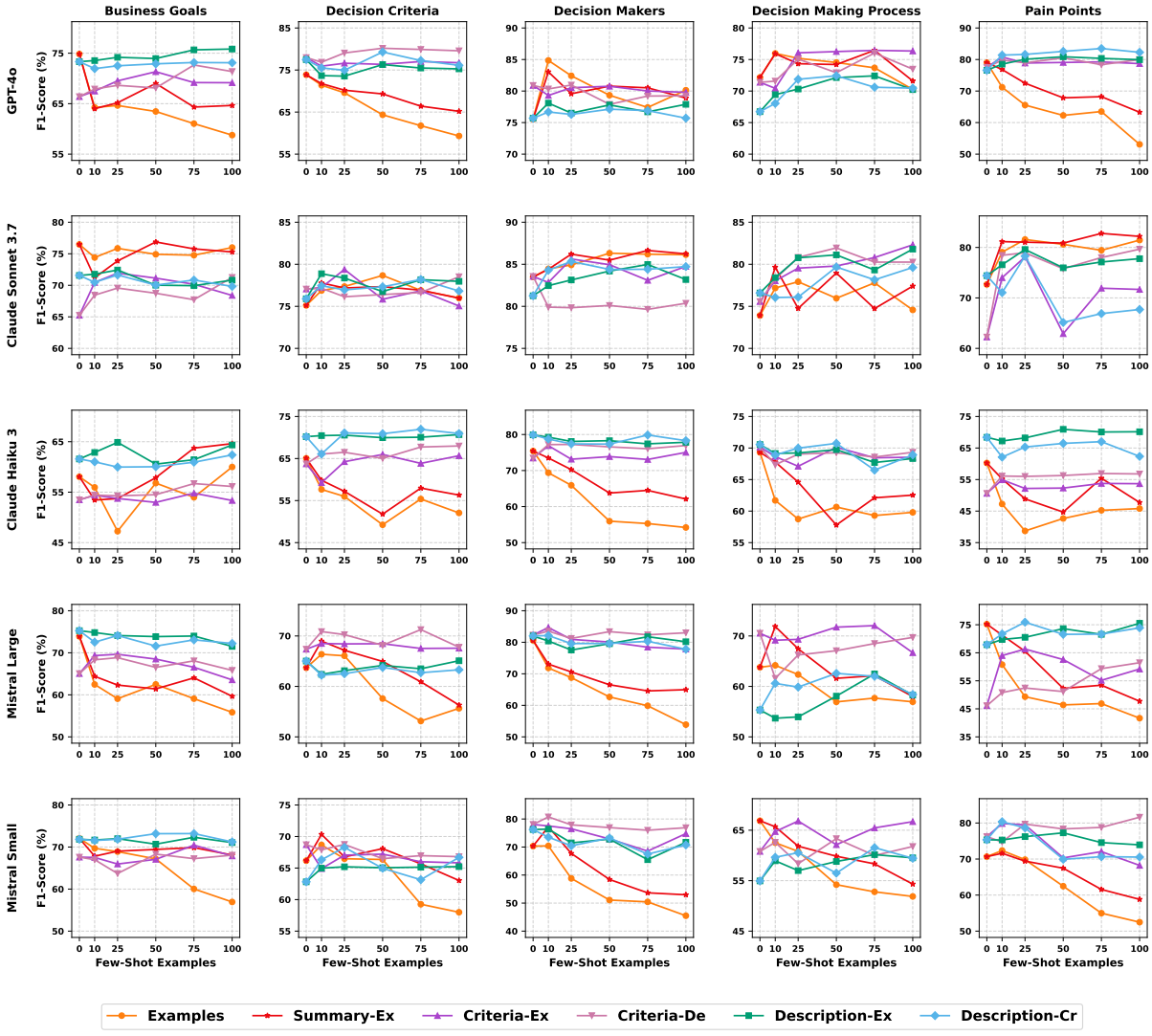

Figure 15: Macro-average F1 performance for all concepts and models.

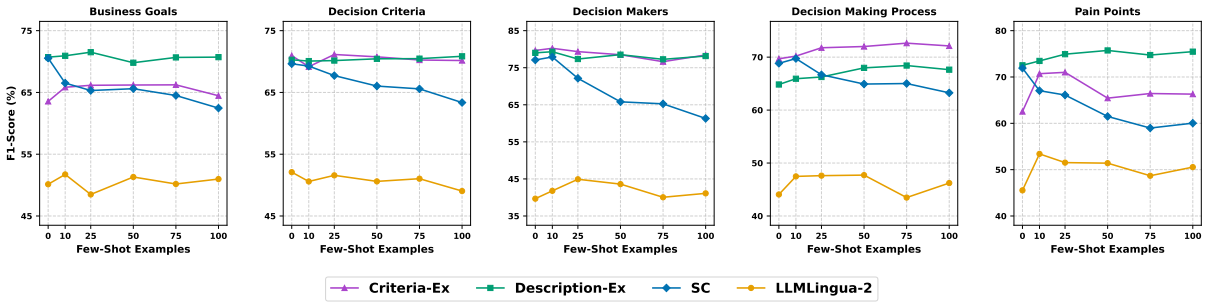

Figure 16: Macro-average F1 performance comparison with token compression methods, averaged over all models.