Transforming the Voice of the Customer by Automating Customer Need Abstraction

by

Artem Timoshenko, Chengfeng Mao, and John R. Hauser

June 2026

Artem Timoshenko is an Associate Professor of Marketing at Kellogg School of Management, Northwestern University, 2211 Campus Drive, Suite 5391, Evanston, IL 60208, (617) 803-5630, artem.timoshenko@northwestern.edu.

Chengfeng Mao is a PhD Student at the MIT Management School, Massachusetts Institute of Technology, E62-366, 77 Massachusetts Avenue, Cambridge, MA 02139, (217) 281-2220, maoc@mit.edu.

John R. Hauser is the Kirin Professor of Marketing, MIT Management School, Massachusetts Institute of Technology, E62-538, 77 Massachusetts Avenue, Cambridge, MA 02139, (617) 253-2929, hauser@mit.edu.

We thank Carmel Dibner, Kristyn Corrigan, John Mitchell, and Maggie Hamilton at Applied Marketing Science, Inc. for insightful discussions and research collaboration. We thank Shitong Xiao for outstanding research assistance. This paper has benefited from presentations at the 2024 Emory Marketing Camp, 2024 Symposium on AI in Marketing, 2024 Insights Association Annual Conference, 39th ISMS Marketing Science Conference, 2025 AIML Conference, Wharton's 3rd Annual Business & GenAI Conference, 2nd Open and User Innovation Conference, 2026 MSI Forum, and research seminars at University of Texas at Dallas, Colorado University at Boulder, and the University of Florida.

Transforming the Voice of the Customer by Automating Customer Need Abstraction

Abstract

Identifying customer needs (CNs) is fundamental to product innovation and marketing strategy. For over thirty years, Voice-of-the-Customer (VOC) applications have relied on professional analysts to manually transform qualitative data into "jobs to be done." This manual abstraction is cognitively demanding, time-consuming, and hard to scale. While current practice uses machine learning to screen content, formulating precise CNs still relies on expert human judgment. We investigate whether CN abstraction can be automated and demonstrate that success depends on how the system is built. In blind professional evaluations across multiple product categories, general-purpose LLMs fall well short. However, a small, fine-tuned model we call a *CN machine* abstracts CNs at least as well as professional analysts, achieving the specificity and source grounding required for industry applications. Supervised fine-tuning does not teach the model to memorize CNs or to reason about customers. Instead, it teaches the syntactic and semantic conventions necessary to abstract raw input into professional-grade CNs. This mechanism holds across foundational architectures and at parameter counts small enough for local, secure execution on modest hardware. By processing entire corpora rather than samples, the CN machine makes VOC studies more comprehensive and surfaces high-leverage insights at scale. Ultimately, the CN machine transforms the allocation of human effort, enabling professional analysts to focus on higher-value-added tasks necessary for successful product innovation.

Keywords: Marketing Research, Product Development, Innovation, Machine Learning, Large Language Models, Consumer Insights

1. Introduction

Understanding customer needs (CNs) is essential for effective product development, product management, and marketing strategy. For example, a snowplow company recognized that customers often need to navigate tight turns between narrow sidewalks. This insight motivated them to develop a zero-turn snowplow brand that immediately solved an important CN. The movie theater industry was revolutionized with stadium seating to fulfill the CNs of a clean and unobstructed view, room to rest arms, easy to get in and out, and storage for refreshments. Identifying patient and physician CNs, such as easy-to-interpret diagnostic information and convenient-sized output, led to a breakthrough medical device (Hauser 1993).

CNs are natural language statements that reveal desired underlying benefits, "jobs to be done." They are articulated at an abstract level to indicate what the product seeks to accomplish, rather than which specific attributes fulfill that need. For example, a clean, unobstructed view in a movie theatre, rather than a 26" wide, swivel chair with cupholders positioned three feet from the next row. Breadth is important; missing important CNs often means the difference between success and failure. Professional studies aim to exhaustively capture CNs that are prioritized using customer surveys and organized in "affinity diagrams" to help product managers focus (Griffin and Hauser 1993).

Identifying CNs is a time-consuming and cognitively demanding manual task. Because customers rarely articulate CNs explicitly, professional analysts must interpret raw qualitative data, such as interview transcripts and online reviews, to distill precise, nuanced insights. This manual approach is expensive, difficult to scale, and slows the time-to-market. However, automating this process is challenging due to the abstract and context-dependent nature of CNs. While keyword searches and machine learning help to screen large datasets, the critical final step of formulating precise CNs still relies on human expertise (Timoshenko and Hauser 2019). If this tedious task were automated, Voice-of-the-Customer (VOC) studies would become faster and more comprehensive, freeing analysts to focus more time on improvements to products and marketing.

Large language models (LLMs) show promise, yet their ability to formulate CNs for professional applications remains an open question. LLMs write text well and have achieved success in abstractive summarization tasks (e.g., Arora, Chakraborty, and Nishimura 2025). However, industry adoption requires precision in capturing the CNs. Poorly formulated CNs, even if automated, would be counter-productive because they focus firms on the wrong sets of questions and do not provide the nuance of "jobs to be done." With prompt engineering, LLMs can generate CNs that "sound reasonable," but these outputs often lack depth, nuance, and reliability (see related research by Gao et al. 2024).

Evaluation is equally challenging. For professional applications, we must assess abstracted CNs with respect to deep insights rather than superficial language. For example, the wood-stain-product CN of "able to achieve a desired finish (e.g., satin, semi-gloss, gloss)" might be expressed by customers as "a product that can give my wood an aged look," "assured the final result is not cloudy," or "can achieve a glossy or flat finish, depending on my preference." Whether or not these phrases represent well-articulated CNs, provide the necessary specificity, and are true to the source material requires professional judgment. Ground truth is hard to obtain, in part because LLMs are trained to provide statements that sound right to the untrained ear (Ji 2024, Selinger 2024).

This paper addresses a question of substantial practical importance in product innovation squarely at the interface of information systems and marketing: can a purpose-built AI system reliably automate the cognitively demanding task of formulating customer needs? Working with an industry partner with over thirty-five years of VOC experience, we obtained professionally formulated CNs to train our model and conducted blinded professional evaluations across five product and service categories. We show that general-purpose LLMs warrant the skepticism practitioners hold; current foundational models formulate CNs well below the professional standards. However, this barrier can be overcome by fine-tuning a small, specialized model that we call a *CN machine*. The CN machine is focused on a single task: abstracting CNs at least as well as professional analysts, capturing the necessary nuance, specificity, and grounding in source material without hallucination. Furthermore, this performance replicates across foundational architectures (including Vicuna, Qwen, and DeepSeek) and outperforms advanced few-shot prompting.

More telling than the fact that it works is *why* it works. Looking "under the hood," we show that fine-tuning does not teach the model to memorize CNs or to reason about customers. Instead, the CN machine is best understood as a dedicated transformation engine: the model learns to recognize what constitutes a valid need – distinguishing true CNs from solutions and preferences – and adopts the syntactic and semantic conventions that turn raw source material into professional-grade "jobs to be done." The task of formulating CNs, at its core, requires short-form abstractive transformations governed by learnable but tacit professional conventions. Experts reliably recognize a poorly formulated or well-formed transformation yet cannot explicitly describe a complete rule for what makes one correct. This is an instance of Polanyi's paradox, that "we can know more than we can tell" (Polanyi 2009; Autor 2014), and learning from many professional examples is precisely the route by which machine learning circumvents this paradox. Tacit conventions are inferred from data rather than from stated rules (Brynjolfsson and Mitchell 2017).

By producing CNs at a consistent level of abstraction and in a consistent form, the CN machine fundamentally reshapes the VOC workflow. Professional analysts are enhanced, not replaced. They can now winnow and prioritize machine-generated CNs directly, bypassing the tedious steps of screening raw source material and abstraction. This single change unlocks scale: VOC studies can be run over far larger corpora, surfacing high-leverage and niche insights that sampling would otherwise miss. Freed from the labor-intensive abstraction phase, analysts refocus on higher-value work such as constructing affinity diagrams and advising clients on strategy. Crucially, this added breadth does not come at the cost of nuance. We show that the machine's CNs remain diverse and preserve niche distinctions rather than collapsing into a few generic statements. The automated system does not degrade to a standardized set of CNs.

Finally, our findings translate into two design principles that we expect to hold beyond this single application. The first is *specialization over scale*: capability for this task comes from task-specific fine-tuning rather than raw model size or frontier reasoning. A CN machine with as few as 0.6 billion parameters matches one more than twenty times larger, and performance plateaus at roughly 600 training

examples. The second is that the model is *built to be governed*: because it is small, it can be fine-tuned and run locally, on a single workstation, without sending proprietary customer data to external services and without exposure to the silent model updates that destabilize prompt-based solutions. Even if a frontier model matched the quality of the CNs, these two features of professional VOC work would still favor a small model that can be owned and run in-house. These design principles extend beyond a one-off application and contribute knowledge upon which other researchers and firms can reuse and build (Hevner et al. 2004; Gregor and Hevner 2013).

2. Related Literature

2.1. Identifying Customer Needs (CNs)

CNs are the basis of the voice-of-the-customer research (VOC, Griffin and Hauser 1993; hereafter, GH 1993). Despite decades of methodological advances in qualitative elicitation, including interview techniques, ethnographic methods, and metaphor elicitation (e.g., Brown and Eisenhardt 1995, Burchill and Brodie 1997, Arnould, Cayla, and Beers 2014, Mitchell 2016, Zaltman 1997), interpretation still relies on human judgment.

As firms recognized that user-generated content (UGC, e.g., online reviews, blogs, and forums) augments customer interviews, new methods emerged to scale CN analysis. While early approaches used topic modeling to identify "bags of words," these failed to capture nuanced CNs (Lee and Bradlow 2011, Lee et al. 2025, Netzer et al. 2012, Schweidel and Moe 2014). To help identify more-nuanced CNs, Timoshenko and Hauser (2019; hereafter, TH 2019) used machine learning to screen diverse informative sentences to be reviewed by professional analysts to abstract CNs. Our paper investigates whether LLMs can successfully automate this final human-centric step.

2.2. LLMs in Business Research

LLMs have demonstrated remarkable capabilities across domains, such as text summarization (Ibrahim 2025), education (Lo 2023), healthcare (Moor et al. 2023, Thirunavukarasu et al. 2023, van Veen et al.

2024), coding (Gao et al. 2023), financial services (Yang et al. 2020), and law (Deroy et al. 2024, Katz et al. 2024). Many of these applications focus on extractive summarization –reproducing existing text, such as pulling product attributes from a review – rather than abstractive transformation. Abstracting a customer need is fundamentally different: it requires generating new phrasing to articulate the implied "job to be done." For example, if a review praises a wood stain because "*it turns pink so I can see where I already applied it*," the model must look past the literal feature to abstract the underlying need: *being able to easily see covered surface areas*.

Within marketing, LLMs have been used in two broad ways. The first treats the LLM as a stand-in for the customer: synthetic respondents that answer survey, conjoint, or elicitation questions in place of real people (Horton 2023; Arora, Chakraborty, and Nishimura 2025; Brand, Israeli, and Ngwe 2023; Qiu, Singh, and Srinivasan 2023). The second treats the LLM as a tool for analyzing data that real customers actually produced, such as recovering decision rules from elicitation transcripts (Dong 2024) or constructing perceptual maps (Li et al. 2024). Our work extends the latter tradition. Rather than prompting a general-purpose model, we build and rigorously evaluate a specialized system to reliably automate the critical professional task of abstracting customer needs.

LLMs do well on many tasks, but not all tasks. Gao et al. (2024, p. 2) suggest that many LLMs "exhibit unstable behavior that differs from human behavior to a statistically significant degree, regardless of the approach used." This variation is particularly prominent in new tasks that the LLM has not memorized from its vast training data. While prompt engineering is effective in some cases (Brown 2020), generating consistently high-quality prompts is challenging (Min et al. 2022; Lu et al. 2022), and wrapper applications often destabilize when foundational LLMs update (Chen, Zaharia, and Zou 2024). Consequently, Challapally et al. (2025) suggest that 95% of industry applications do not achieve their targets.

These limitations reflect on a broader theme in information systems (IS) research: an AI system's value often derives less from raw model capability than from how the system is designed, evaluated, and embedded in expert work. Indeed, our results show that capability is not the binding constraint, as highly

capable foundational LLMs fail at CN abstraction while small fine-tuned models succeed. Our work relates to two central IS strands: evaluating professional standards and redividing labor. Regarding evaluation, AI tools are easily mis-evaluated against static "ground truth" because much professional competence is tacit know-how that static labels do not capture (Lebovitz, Levina, and Lifshitz-Assaf 2021). Therefore, rather than scoring CNs against a fixed rubric, we rely on blinded professional judgments (van der Lee et al. 2019). For the division of labor, embedding AI tends to redivide rather than replace human tasks, yielding hybrid human–machine workflows (van den Broek, Sergeeva, and Huysman 2021; Baird and Maruping 2021). We demonstrate how a narrow, well-specified system automates a task previously reliant on human judgment, enabling a new division of labor in VOC studies.

3. Industry Practice

Before asking whether an automated system can abstract CNs, we must understand what professional analysts actually do in the VOC studies. This section describes standard practice and distills it into the requirements we use to evaluate the CN machine.

Formulating CNs is demanding. The challenge lies in understanding the deeper motivations that drive customer behavior and capturing them in a concise form. Industry professionals often differentiate between CNs, solutions, targets, and opinions. For example, a customer might express dissatisfaction with cellphone battery life (an opinion), but the underlying CN might be the desire for longer, uninterrupted use while traveling. In academic literature, solutions and targets are often framed as product attributes, while opinions reflect customer sentiments.

Understanding well-formulated CNs before focusing on specific solutions provides insights beyond current market offerings. Before the zero-turn snowplow was launched, the ability to move between perpendicular sidewalks was not a defined attribute. Similarly, prior to stadium seating in movie theaters, attributes such as elevation and cupholders were not part of the conversation. In services, a professional organization may wish to attract members, but not recognize that new members must learn the specialized informal jargon before they can benefit from conferences, training, and certification.

To identify CNs, firms use experiential interviews, metaphor elicitation, ethnography, focus groups, call center logs, or user-generated content to create a corpus of phrases. Professional analysts, with extensive training and experience, highlight relevant sentences in the source material and interpret the customers' words as CNs, keeping the verbiage and intentions as close to the original as feasible. Because customers rarely articulate CNs clearly, the analysts' task involves interpretation (abstraction) rather than mere paraphrasing.

Example 1. A customer complained about a 30-second timer in a toothbrush: "*I replaced an old brush with a new one, BUT the description doesn't say that this model no longer has a 30-second timer. The brush shuts off after 2 minutes but the 30 second timer is missing. I would not have purchased this product if I had known.*" From this review, a professional analyst abstracted a CN "*Able to know the right amount of time to spend on each step of my oral care routine.*"

CNs must be nuanced, yet not so specific that they forestall creativity. Overly generic CNs, such as "*ease of use*," fail to provide actionable insights and meaningful ideas for innovation. Conversely, too-specific CNs, such as "*able to dry in 20 minutes at 70% humidity*" limit creativity and fail to generalize.

Example 2. A professional study identified three CNs focused on breath freshness: "*Able to eat and drink anything and my breath still stays fresh*," "*Able to have fresh breath all day, i.e., no need to keep freshening it*," and "*Able to tell if I have bad breath*." Each CN identifies a specific "job to be done" related to a more-general area of breath freshness. If the product developer or marketer identifies ways to satisfy these CNs, customers might pay a premium for the solution(s).

Abstracting CNs is a cognitively demanding task, and humans are fallible. GH 1993 report that each analyst was able to identify only 54% of the CNs (range 45-68%) that were ultimately identified. With more applications, professional analysts have become more skilled, but not perfect. New analysts receive training materials that define CNs, contrast CNs with (existing) solutions, and provide standards to abstract CNs. They learn best practices such as formulating CNs as concise positive statements using simple, accessible language that captures the core customer benefit without ambiguity. Analysts are given

many examples of CNs, including statements that are not CNs. The analysts hone their craft by formulating CNs and receiving feedback from more experienced colleagues. We attempt to replicate the spirit of this training with example-based fine-tuning.

Depending on the product (or service) category and the source material, analysts might identify hundreds or thousands of potential CNs, creating significant redundancy. Using keyword matching and experience-based judgment, analysts "winnow" the identified needs to a less-redundant set, typically 80-120 CNs. To focus product development, product management, and marketing on creative solutions, they organize these into a three-level hierarchy of primary, secondary, and tertiary CNs (Burchill and Brody 1997, GH 1993). In our paper, the term "customer needs" refers to the highly nuanced tertiary needs in the hierarchy – the critical step in the VOC process. An important criterion for the effectiveness of VOC studies is its breadth: the ability to identify tertiary CNs across every primary and secondary branch of the hierarchy.

Taken together, professional practice defines what a successful automated solution must do. Each abstracted CN must read as a genuine customer need rather than a solution, target, or opinion; it must be specific enough to guide innovation yet general enough to preserve creative latitude; and it must follow from the source material without fabrication. Beyond the individual statement, the full set of CNs must achieve breadth, populating every primary and secondary branch of the affinity hierarchy. These requirements are not ours to invent. They are the standards professional analysts are trained to meet, and they are the standards against which we evaluate the CN machine in the studies that follow.

4. Large Language Models for Extracting Customer Needs

To examine whether LLMs can formulate nuanced CNs, we developed prompts for a foundational LLM and gathered training data for fine-tuning the model (SFT LLM). We then recruited blinded professional analysts to evaluate the CNs produced by the LLMs and by human analysts in the original VOC application.

4.1. Foundational Model: Vicuna 13B

Our primary analysis uses Vicuna 13B as a foundational LLM. Our preliminary investigation showed that CNs formulated by Vicuna 13B were qualitatively similar to those from contemporary state-of-the-art models like ChatGPT-4o, providing a strong baseline. We observed no noticeable improvement with the larger Vicuna 33B.

Vicuna is a general-purpose LLM, developed by researchers from UC Berkeley, UCSD, and CMU (Chiang et al. 2023). It is built on LLaMA 2 and is fine-tuned with 70,000 ChatGPT conversations (Touvron et al. 2023). Vicuna performs comparably to the contemporary open-source and closed-source LLMs in writing tasks (Zheng et al. 2024, Zhang et al. 2023) and has been applied successfully in downstream applications and research (Zhu et al. 2023, Mullappilly 2023).

For our baseline, we explored multiple prompt variations, starting with "*Extract customer needs from <Review Text>*," and then adding (1) a definition of a CN, (2) a requirement to formulate a response in a single sentence, (3) examples of CNs (in-context learning), and many other variations and combinations. In exploratory preliminary analysis, all options performed similarly. Our primary analysis uses the following prompt: "*For a <Product Category>, identify a customer need from the user review. If no need is found, return []. Review: <Review Text>*".

Our primary empirical evidence uses blind studies with professional analysts to evaluate the Vicuna-based CN formulations. We additionally explore alternative foundational LLMs and prompting approaches using a trained classifier (Gu et al. 2024; Zhang et al. 2025).

4.2. Fine-tuning an LLM with Professional VOC Studies

We fine-tuned the LLM to abstract CNs using examples from professional VOC studies. Fine-tuning improves the LLM's performance and output style by directly updating weights of the foundational model (Dong et al. 2023). This approach mimics the training of professional analysts, who master the CN

abstraction task through exposure to examples rather than following explicit rules.[1]

Our fine-tuning data span ten different product categories, such as activewear, glucose monitors, and recreational vehicles. The data came from professional VOC studies conducted by a firm with over 35 years of experience and hundreds of successful applications in product development, product management, and marketing. Online Appendix A provides a complete list of product categories. For each category, the firm's analysts reviewed interview transcripts and/or online reviews to identify CNs. Crucially, the firm retained the original source material or matched each CN to the corresponding source sentence. Our data include these matched verbatims. Because firms typically lack the incentive to meticulously retain these matched verbatims, our dataset is unique.

The ten professional VOC studies provided 1,549 CN-verbatim pairs. We randomly split the "positive" examples into two subsamples for model training (80%) and validation (20%). We add a small sample of uninformative "negative" examples so the model learns to return '[]' when the input contains no CN. All evaluations of the LLMs in this paper use out-of-sample product categories excluded from fine-tuning.

Online Appendix B illustrates positive and negative examples from our training data. The question-and-answer structure is similar to instruction fine-tuning (Chung et al. 2024). The data provide the LLM with examples of ideal answers to application-specific prompts. We add a special tag <GPT-VOC> and indicate the product category to condition the model for the (new) CN abstraction task (Schick et al. 2024; Shen et al. 2024). After fine-tuning, this standardized prompt acts as a shorthand instructing the SFT LLM that the task is to abstract CNs.

4.3. Illustrative Example of LLM-Based CNs

Figure 1 illustrates the output for out-of-sample wood stain products. From an online review, professional analysts identified the CN: "*Able to see what surface areas I have already covered*." This CN seems

[1] Fine-tuning was facilitated by the DeepSeek Library and required approximately eight hours using four Nvidia A100 GPUs from Lambda Labs – an on-demand AI developer (GPU) cloud service. After calibration, applications can run on a local workstation.

relevant when wood stain is applied to larger surface areas.

Figure 1. Illustrative Example Customer Needs Identified from Online Reviews

Without fine-tuning, the foundational LLM abstracted a CN "*Easy to see coverage*." While this statement correctly summarizes the topic of the online review, it lacks the specificity required to guide innovation. The performance of the foundational LLM in Figure 1 is typical and is relevant for other foundational LLMs – a typical failure mode where foundational models merely paraphrase reviews or elicit solutions.

Conversely, the SFT LLM formulated: "*Assured that I can see where I have applied the stain*." This CN captures a "job to be done" from the original content, is concise, and provides sufficient detail for product development. The formulation by the SFT LLM includes a clarification ("*e.g., it turns pink and is visible*"). Our training data contain similar clarifications in 34% of the examples, and the SFT LLM picked that up. Online Appendix D provides additional examples of online reviews and the corresponding CNs as abstracted by professional analysts and LLMs. While the SFT LLM formulations are different from professional analysts, they capture similar meaning and adhere to similar professional standards – a judgment we examine formally in the following sections.

5. Empirical Evaluation: Study Overview

We evaluate the LLMs' ability to abstract CNs using five interrelated studies (Table 1). The studies vary in research questions, data sources, and industries. The methodologies in each study are matched to the

specific issues and the available data. When issues overlap, the variation in methodologies ensures that the resolution of the underlying issue is not methodology dependent.

Table 1. Study Overview

Study	Product Category	Primary Research Question(s)	Method & Data Source
Study 1	Wood Stain Products	Do SFT LLMs abstract customer needs as well as professionals?	Blind evaluation by professional experts; online reviews and professional forums.
Study 2	Wood Stain Products	Does the foundational LLM impact performance? How much training data are necessary?	Sensitivity analysis using fine-tuned transformer-based quality classifiers.
Study 3	Wood Stain Products	Does SFT facilitate transformation or memorization? How do failure modes differ from foundational models?	Comparative failure-mode analysis and cross-evaluation via performance probes.
Study 4	Oral Care Products	Do results generalize to other product categories and different data sources?	Blind evaluation by professional experts; Amazon reviews and CNs from prior research.
Study 5	PDMA (Service Category)	Do insights extend to services and experiential interviews?	Professional VOC application of the SFT LLM in the service industry. Short descriptions of four other applications.

Our primary evaluations (Studies 1, 2, and 3) focus on wood stain products. Study 1 provides the core evidence comparing LLM-formulated CNs to the professionally formulated ones. Study 2 explores the limits of the CN machine by evaluating different foundational models, model sizes, and amounts of training data. Study 3 investigates what LLMs learn during fine-tuning. We contrast the transformation and memorization mechanisms, and cross-evaluate failure modes for the foundational and SFT models.

Studies 4 & 5 provide evidence from other industries to establish cross-category generalizability. Study 4 reports an additional blind study with professional analysts in the oral care product category (toothbrushes/toothpaste), providing a robust replication of our primary findings. Study 5 evaluates CNs abstracted by the SFT LLM in a service category using interview-based data, highlighting rich managerial insights and resulting improvements in the service. None of the three categories—wood stain, oral care, and professional services—were included in the fine-tuning data.

Evaluating the abstracted CNs requires domain experts to judge whether the statements adhere to

professional standards (Griffin 2004; PDMA's Glossary for New Product Development). Our research partner uses extensive training and peer support to ensure that the definition of CNs is consistent with professional standards shared across the discipline, rather than a single firm's house style. Besides conducting numerous VOC studies for consumer brands and business-to-business organizations, our research partner is called upon to train other firms in CN identification. From this partner, we recruited professional evaluators with expertise comparable to that of the original analysts but who were not involved in the initial for-client studies for the evaluated product categories. The evaluators were blind to whether the CNs were formulated by human analysts, the foundational LLM, or the SFT LLM.

6. Studies 1-3: Professional Analysts vs. LLMs for Wood Stain Products

The empirical analysis in this section builds on a professional VOC study that identified 103 CNs for wood stain products. We augment these data with LLM-based CNs. Specifically, the original VOC study used a machine-learning approach to screen 14,341 online-review sentences and identify 1,000 informative and non-repetitive sentences to ensure diverse content (TH 2019). Following standard procedures, the professional analysts read the selected online reviews and manually identified unique CNs. The firm shared the source verbatims for every unique CN, allowing us to apply the LLMs to identify CNs from the same texts.

To minimize information leakage, the wood stain category was excluded from LLM fine-tuning. Although some information on wood stains might have been available during the foundational LLM training, professional VOC studies and CN formulations are trade secrets and unlikely to be available publicly. Regardless, any leakage would only reinforce the key qualitative findings in this section.

6.1. Study 1a. Are Statements Abstracted by LLMs Well-Formulated Customer Needs?

Our first study evaluates whether CNs abstracted by professional analysts and LLMs adhere to the industry's professional standards. Three professional evaluators, who were not involved in the original wood stain application but possess extensive VOC experience, evaluated CN statements on three

dimensions. The wording of the questions to the evaluators was based on extensive discussion with the firm's VOC experts, and the questionnaire was pretested to ensure clarity and validity. Online Appendix E provides the exact instructions. The basic questions asked are paraphrased below.

(1) The statement qualifies as a CN identified in a typical VOC study ("Is Customer Need")

(2) The statement captures sufficient detail about a CN ("Sufficient Detail")

(3) The statement is based on some information in the review ("Follows from the Source")

Each blind evaluator assessed 150 randomly-chosen sentences from online reviews. For each sentence, an evaluator was given the text and three randomized CN statements (professional analyst, foundational LLM, and SFT LLM). Evaluators were blind to the purpose of the study, and posttests indicated no inadvertent cues about how the CNs were extracted. We aggregated individual evaluations using majority voting. In the following analysis, each data point corresponds to a sentence-CN combination.

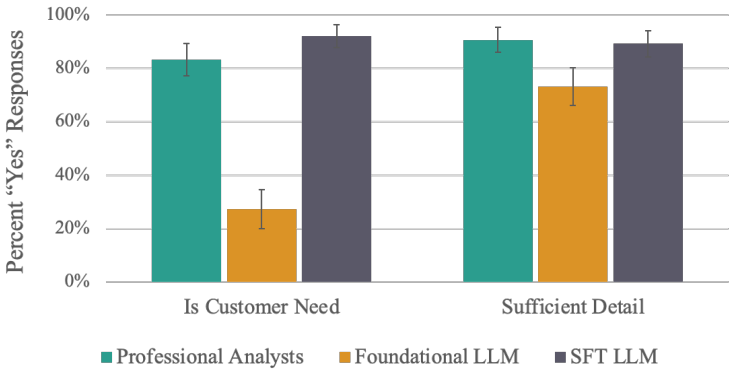
The sample of 150 review sentences included (1) 90 *verbatims* leading to CNs in the VOC application, (2) 30 sentences indicated by the firm as *informative* but not used as verbatims, and (3) 30 sentences indicated as *uninformative* (rejected as lacking CNs in the original study). For *verbatims*, all three approaches identified CNs. For the other categories, where original analysts did not formulate CNs, we augment LLM outputs with randomly selected analyst-identified CNs from the VOC application to maintain blinding and avoid inadvertent signals.[2]

Figure 2 reports results for the first two questions, aggregated across verbatim, informative, and uninformative reviews. (Online Appendix F reports the disaggregated results and interpretations. The interpretations are consistent with the main text.) The professional-analyst CNs provide an important benchmark. Professional analysts identified these CNs in a for-client application. Despite extensive training, professional analysts are not perfect; in 1993, Griffin and Hauser reported only 54% capture of

[2] After considering many alternatives ways of choosing analyst-identified CNs, we settled on randomly chosen, but real, CNs. This strategy is faithful to the original VOC professional study and consistent with the reality that human analysts do not identify 100% of the CNs.

CNs from experiential interviews. In Study 1, the professional evaluators agreed with the original analysts about 80% of the time on "Is Customer Need" and "Sufficient Detail." Although the task differs from GH 1993, the 80% agreement suggests improvement in industry practice and reinforces our decision to rely on professional evaluators rather than research assistants or untrained crowdsourced workers.

Figure 2. Comparison of Customer-Need Abstraction by Professional Analysis and LLMs



We use a z-test to test the statistical significance of differences between models. The patterns with McNemar's paired-proportions test are consistent. We use Cohen's h to report effect sizes. The error bars in Figures 2-8 represent 95% confidence intervals.

SFT training is effective. The SFT LLM CNs are significantly more likely to be judged as typical-to-VOC CNs than those identified by analysts, though the effect size is small ($z = 2.29$, $p = 0.02$, $h = 0.27$). On Sufficient Detail, the SFT LLM CNs are not significantly different from the analysts, slightly favoring the SFT LLM ($z = 0.40, p = 0.69$, $h = 0.04$). These results validate the use of SFT LLMs to identify CNs.

Figure 2 indicates that the foundational LLM is not a viable solution for this product category. It performs significantly and substantially worse than both professional analysts ($z = 9.76$, $p < 0.01$, $h = 1.20$) and an SFT LLM ($z = 11.42$, $p < 0.01$, $h = 1.47$). It also performs significantly worse than at capturing detail, although the effect sizes are moderate ($z = 3.91, p < 0.01, h = 0.47$, and $z = 3.56, p < 0.01, h = 0.42$, respectively).

We next examine whether hallucinations are a problem for LLMs or analysts (Rawte et al. 2023).

Because customers do not state CNs directly, both must interpret information to abstract the implied "job to be done." Analysts might hallucinate after tediously reading many source sentences. LLMs might hallucinate based on the extensive data in foundational training.

In Figure 3, we evaluate whether the abstracted CNs follow from the *verbatims* known to contain a CN. The evaluators answered: "Please evaluate whether or not the statement is based on information in the review. In particular, is it reasonable that a VOC study would abstract this customer need from the review?"

Figure 3. CNs Identified by Analysts and LLMs Capture Information from the Reviews

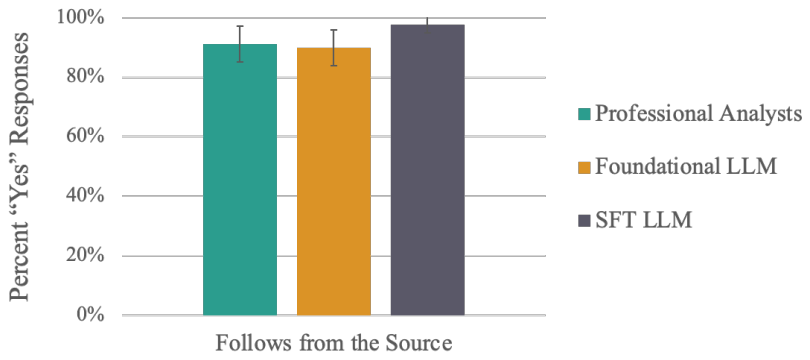


Figure 3 focuses on the 90 *verbatims.* Limiting Figure 3 to verbatims provides a fair comparison among professional analysts and LLMs and, if anything, favors professional analysts. Online Appendix F reports data from all other review sentences; the data are face valid and reinforce our interpretations.

The professional-analyst-abstracted CNs were judged to closely represent information from the verbatims for 91% of observations. The firm reports that this percentage is within an expected range for VOC applications. The SFT LLM significantly outperforms professional analysts although the effect size is modest ($z = 2.53, p = 0.01, h = 0.31$). The foundational LLM performs similarly to the professional analysts ($z = 0.33, p = 0.75, h = 0.04$). Figure 3 suggests that LLM hallucinations are not a practical problem for identifying CNs.

6.2. Study 1b. Can LLMs Capture Diverse Customer Needs?

Typical VOC applications identify hundreds of CNs from raw qualitative data. To be useful for consumer

insights, CNs are winnowed to remove duplicates and grouped into an affinity diagram—a hierarchical structure of primary, secondary, and tertiary CNs (Burchill and Brody 1997; GH 1993). Higher-level primary and secondary CNs are chosen to represent broad customer perspectives. For example, the tertiary wood stain CN "a*ble to control depth and finish of the stain and topcoat*" could belong to the secondary CN "*application looks even and consistent,*" under the primary CN "*appearance and finish*." Professional experience suggests that applications are most successful when a VOC study identifies CNs in every primary and secondary category. To evaluate whether the LLMs capture this breadth, we use an affinity diagram based on a concatenation of the professional-analyst and SFT-LLM CNs. Our research partner developed this diagram in three steps.

Step 1. Preliminary Winnowing: Winnowing relies on human judgment to eliminate redundancy, often merging CN-statements at the tertiary level. Our research partner first winnowed the SFT-LLM-identified CNs obtained from the full corpus of 14,341 source verbatims. Analysts used standard professional procedures to return a preliminary set of 154 winnowed CNs.

Step 2. Final Winnowing and Affinitization: We merged the 154 winnowed SFT-LLM-identified CNs with the 103 analyst-identified CNs from the original VOC application. Experienced analysts, not involved in the original application or in Study 1, constructed an affinity diagram. Following standard practice, they further winnowed the combined pool to a final set of 117 CNs. The analysts were blind to the source of the CNs and preserved the mapping from the 257 (154 plus 103) initial CNs to the final 117.

Step 3. Mapping CNs Back to Verbatims: Our research partner reconstructed the mapping from a randomly-sampled set of 2,000 pre-winnowing SFT-LLM-identified CNs to the final 117 CNs. Professional analysts performed this many-to-many mapping manually.

This process highlights a fundamental workflow shift. In standard practice, analysts winnow the raw source material prior to CN abstraction because abstracting CNs from every potential verbatim would be prohibitively time-consuming. However, winnowing raw text is substantially harder and more time-consuming than winnowing abstracted CNs. The SFT LLM flips this constraint. Because automated abstraction is virtually costless, professional analysts bypass raw text and winnow machine-generated

CNs directly. This efficiency enables VOC to scale to much larger corpora while saving analysts' time.

The mapping required a major (and expensive) effort by our research partner over several months. It would have been cost-prohibitive to reconstruct a mapping for the entire set of 14,341 source verbatims. In typical VOC applications, the mapping from final CNs back to source verbatims is rarely retained because the substantial cost is not justified by the business benefits. Indeed, in over thirty-five years of practice, our research partner recalls only one prior instance requiring such mapping: a litigation-support application demanding extensive documentation.

Strategic Value: Managerially Relevant Primary and Secondary Customer Needs

For wood-stain products, the final affinity diagram includes 30 secondary CNs grouped into eight primary CNs. The CNs identified by the SFT LLM had slightly more strategic breadth, populating 100% of the primary and secondary groups. In contrast, the original VOC study identified CNs from seven (87.5%) of the primary groups and 24 (80%) of the secondary groups. For example, the original analysts missed the primary CN describing *a desire for products that help make the maintenance of wood easier*.

Looking at the tertiary level, the 154 winnowed SFT-LLM-identified CNs account for 84.6% of the final affinitized tertiary CNs, while the 103 analyst-identified CNs account for 48.7%. (The percentages add to >100% due to overlap.) These figures must be interpreted cautiously. First, cost considerations limited the original VOC study to a sample of the corpus, whereas the SFT-LLM processed the full corpus. Second, these percentages confound redundancy reduction with whether a final tertiary CN is equivalent to an abstracted statement.

6.3. Putting Together Studies 1a and 1b

Study 1a evaluates customer needs one at a time, and by construction, it cannot detect a subtler risk: the CN machine might produce needs that are individually plausible yet collectively narrow, converging on a few standardized statements at the expense of diversity and nuance. Whether the machine homogenizes its output is a property of the full set of needs, so we examine the question at the level of the affinity

diagram.

The overlap structure in Study 1b speaks directly to this concern. If the CN machine were flattening its output toward a few generic needs, its statements would concentrate in a handful of categories, leaving much of the hierarchy unpopulated. Instead, the machine-identified CNs populate every primary and secondary branch and preserve nuanced distinctions at the tertiary level. This breadth survives winnowing – a process explicitly designed to remove redundancy. A homogenized set of CNs could not both survive redundancy reduction and still span the hierarchy. Thus the diversity we observe reflects genuine variation in the underlying output rather than sheer volume.

The same comparison shows that the CN machine recovered valid needs that the original study missed. We attribute this to corpus coverage rather than to any sentence-by-sentence advantage over analysts. A manual study samples the source material to control costs, whereas the machine can process the full corpus. The practical consequence is comprehensiveness: CNs that sampling would leave undiscovered, including niche needs relevant to smaller customer segments, are surfaced. Taken together, the CN machine makes VOC studies broader without making them blander.

A clarification about scope is warranted. Our primary foundational baseline is Vicuna. Later models, including the Qwen and DeepSeek families in Study 2, are stronger general-purpose systems. The trajectory of the field makes it reasonable to expect that general-purpose models will continue to improve, potentially abstracting CNs better than Vicuna and approaching the quality of a fine-tuned model with careful prompting. Our claim is therefore deliberately focused. We demonstrate that (1) an untrained foundational model must be rigorously tested before it can be relied upon, and (2) fine-tuning produces a reliable CN machine even when the foundational LLM is not the frontier model. Study 2 examines whether the CN machine concept is valid for small foundational models. These operational insights will likely hold regardless of which foundational model is strongest at the time of use.

6.4. Study 2. Generalizability and Boundary Conditions

We now explore the boundaries of the fine-tuned LLM's performance by testing the impact of training

data size, model size, and generalizability across different foundational models. Additionally, we evaluate whether few-shot prompting bridges the gap between foundational and fine-tuned models.

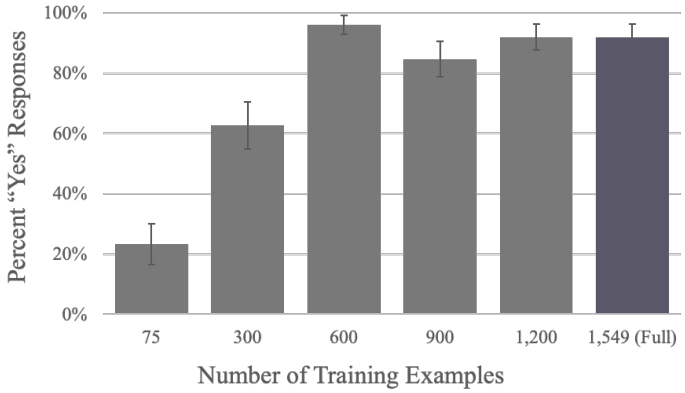
Because high-cost professional-analyst evaluation is infeasible for these many variations, we trained a transformer-based classifier to reproduce the professional evaluators' judgments for each of the three Study 1 questions (henceforth, "the classifier"). Our approach is motivated by recent literature. Research suggests that LLMs are scalable and particularly good at open-ended evaluations, but without additional calibration, may be biased (Shi et al. 2025; Panickssery et al. 2024). For focused yes/no tasks, pretrained transformer-based quality classifiers are particularly effective (Bucher and Martini 2024; Gu et al. 2024; Zhang et al. 2025). Unlike the CN machine which must perform the generative task of abstracting nuanced CNs from source material, the classifier must only perform a binary evaluation of whether an abstracted statement meets professional standards.

We validated the classifier against held-out human evaluations and found no statistically significant differences across the nine metrics. Notably, it accurately predicted performance on SFT LLM outputs even though no SFT LLM examples appeared in its training. This provides a reliable proxy for human judgment at scale. Online Appendix G provides construction details and full validation.

Impact of Training Sample Size on Fine-tuning

Figure 4 reports the classifier-predicted performance for the "Is Customer Need" metric as the number of fine-tuning examples varies. Results for "Sufficient Detail" and "Follows from the Source" are available in Figure H1 in Online Appendix H. We find that performance improves steadily as the training set grows from 75 to 600 examples, but marginal gains diminish beyond this point. Increasing the training set size above 600 examples does not further enhance the SFT LLM's performance. A similar plateau occurs for the other two metrics, where 600 examples are sufficient to achieve near-optimal results.

Figure 4. Predicted Performance with Different Amounts of Fine-tuning Data

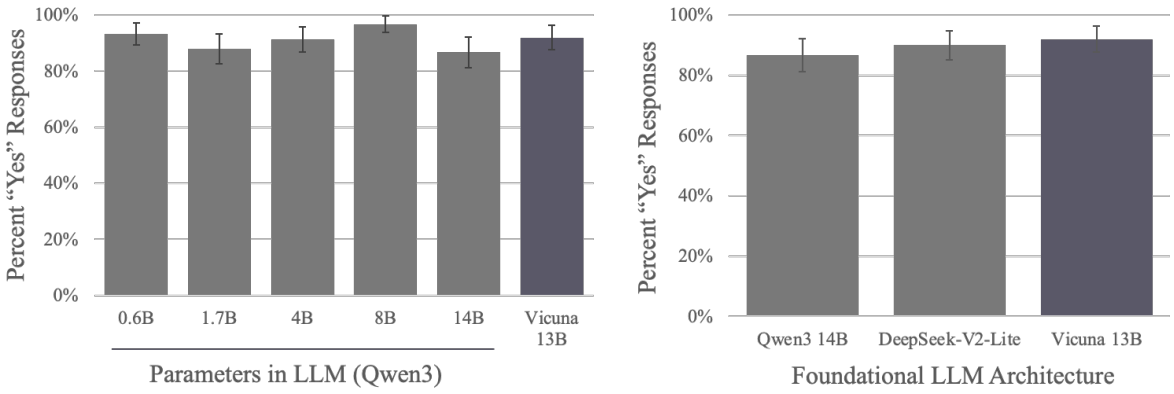


These findings are encouraging for practitioners, particularly for smaller firms with limited resources. While high-quality, professionally abstracted CNs are expensive and difficult to obtain, these results suggest that an SFT LLM can achieve high performance with substantially fewer than 1,549 training pairs. Furthermore, our analysis of category diversity (Figure H2) indicates that the model is robust to the breadth of training data: performance remained consistent whether the 600 examples were drawn from all ten product categories or a subset of only four.

Model Size and Generalization Across Foundational LLMs

Figure 5 reports classifier-predicted performance on "Is Customer Need" as we vary (a) the number of parameters in the foundational LLM and (b) the foundational LLM itself. The results for "Sufficient Detail" and "Follows from the Source" are reported in Figure H3. To evaluate the impact of model size, we fine-tuned the Qwen3 series, an open-weight model family that offers a range of parameter counts with a shared architecture. We focus on the models with 0.6B, 1.7B, 4B, 8B, and 14B parameters. To assess generalizability to different foundational LLMs, we compare Qwen3-14B with DeepSeek-V2-Lite, which represented the state-of-the-art across popular benchmarks at the time of this study. Vicuna-13B is included in both analyses as a baseline.

Figure 5. Predicted Performance of Different Foundational LLM Architectures



The first panel of Figure 5 indicates that increasing the number of parameters in the foundational LLM does not substantially improve performance after fine-tuning. For example, the Qwen3-0.6B model achieves results comparable to the Qwen-14B version and the Vicuna-13B baseline. The second panel suggests that our findings are not unique to Vicuna. Both Qwen and DeepSeek were released one and a half years after Vicuna and, with fine-tuning, appear to perform just as well.

These results provide evidence regarding the fundamental complexity of identifying CNs. On one hand, identifying nuanced CNs is inherently challenging; customers rarely articulate their needs explicitly, so current practice relies on professional analysts with extensive training to capture the underlying "jobs to be done." On the other hand, from a computational perspective, the task involves short-form abstractive summarization of limited text segments, which may not require massive reasoning capacity. Our findings demonstrate empirically that this complex abstraction problem is highly feasible even for small-scale LLMs fine-tuned on relatively modest datasets.

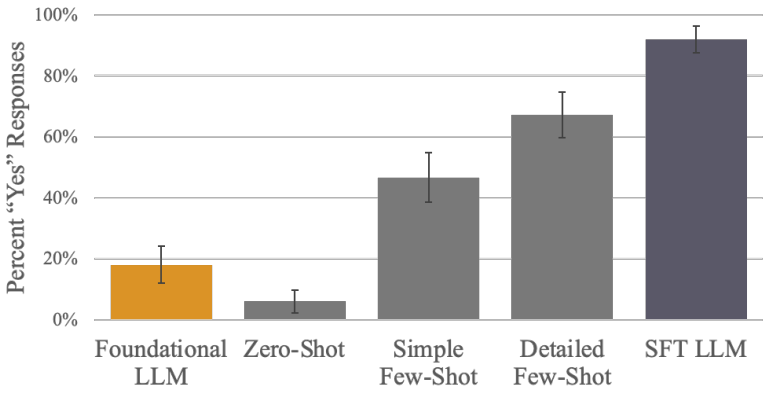
Evaluating Alternative Prompting Strategies

Prior to Study 1, we pretested a wide variety of prompting methods. Using the classifiers, we evaluate these alternatives more formally. Figure 6 reports classifier-predicted performance for the "Is Customer Need" metric as we vary prompting approaches. The zero-shot prompt includes a formal definition of a CN, detailed instructions for formulation (such as requesting concise language and contrasting CNs with solutions), and a defined output format. The simple few-shot prompt removes these detailed instructions

but adds 10 CN examples from various product categories. Finally, the detailed few-shot prompt combines both the extensive instructions and the examples.

For the wood-stain category, none of these prompting strategies is sufficient for “Is Customer Need," although few-shot prompting does well for “Sufficient Detail” and “Follows from the Source” (See Figure H4). While increasing LLM context windows may allow for more examples to further improve few-shot performance, we caution that prompting remains fundamentally susceptible to foundational model updates and model drift (Chen et al. 2024). In contrast, fine-tuning allows LLMs to perform at least as well as professional analysts, providing a stable and reliable solution. This is a critical distinction for practical applications; the primary bottleneck is not the computational cost of SFT, but rather the reliability of the generated customer insights.

Figure 6. Predicted Performance of Alternative Prompting Strategies



6.5. Study 3. Towards a Deeper Understanding of the SFT LLM Performance (Why it works)

Transformation versus memorization

Figures 4 and 5 suggest that CN abstraction is a specialized task requiring moderate fine-tuning data and model size. Furthermore, we found that neither the foundational LLM nor the SFT LLM can articulate deep, nuanced professional-grade CNs when prompted without source material (e.g., simply asking, “*what are customer needs for wood stain products?*”). Together, these observations suggest that the SFT LLM is not storing CNs for retrieval, but is instead learning to transform source material into a specific

format. The SFT LLM does not "learn" the customer's perspective as a knowledge base; it learns a process. While the foundational model likely contains knowledge about wood stain, it cannot leverage that out-of-context knowledge to articulate CNs with sufficient breadth, detail, and semantic/syntactic correctness (Treutlein et al. 2024).

To further examine memorization and transformation, we selected prompts that required the LLM to draw on stored general knowledge. Consider a well-known closing line from Dickens' (1859) A Tale of Two Cities: "*It is a far, far better thing that I do, than I have ever done; it is a far, far better rest that I go to than I have ever known*." This quote and its scholarly analyses are almost certainly present in the training data for the foundational LLMs. Not surprisingly, a foundational LLM prompted for CNs recognizes the context and suggests the CNs of *meaning, purpose, redemption,* and *closure*.

In contrast, the SFT LLM focuses strictly on the task it was trained to do. Ignoring the broader literary context, the SFT LLM returns the abstracted CN: "*a better rest*." This response reflects a task-specific induction: a semantic shift from source to abstracted need, and a syntactic alignment to the preferred format (Ouyang et al., 2022). This pattern holds for other well-known quotes that were likely in foundational training. For example, given Dickens' opening line, "*It was the best of times, it was the worst of times*," the SFT LLM ignores the literary context and returns []. These examples suggest that the CN machine functions as a dedicated text transformer rather than a knowledge retriever.

If the CN machine succeeded by memorizing product-specific phrases from its training data, its output would track the surface vocabulary of the input. It does not. Consider three reviews that voice the same underlying need in very different registers: a formal complaint ("*The temporal longevity of the power cell is insufficient for my professional requirements*"), a slang post ("*Bro, this phone is cooked, the battery dies in like two hours, fr*"), and a technical specification ("*Device exhibits high discharge rates; 5000mAh capacity fails to meet 18-hour duty cycle*"). The three share almost no words, yet the SFT LLM abstracts each into the same professionally formed need: *Assured that the device holds its charge for a long time*, phrasing varies slightly with a few additional words. The CN machine has learned the regularities that turn varied source language into a professional need statement, rather than a cataloging

CNs tied to particular words or products.

Declarative knowledge versus procedural skill

To test whether the CN machine acquires declarative knowledge (the ability to state and reason about abstraction rules) rather than just the procedural skill of applying them, we probe the SFT LLM's understanding of the VOC approach (Table 2). The first question poses a reasoning challenge, which the foundational LLM answers quite well. It correctly parses the prompt and arrives at a valid answer. In contrast, the SFT LLM fails this reasoning test, instead transforming the input into a CN.

Table 2. Questions to Examine the SFT LLM's Understanding of CN Abstraction

The second question examines conceptual understanding of the CN-abstraction task. The foundational LLM can state the rules for CN abstraction. The SFT LLM, which is quite good at CN-abstraction, has "forgotten" the rules for CN abstraction (consistent with Luo et al. 2025). Rather than explaining the rules, the SFT LLM simply executes a semantic and syntactic transformation of the prompt into a valid CN. These diagnostics suggest that fine-tuning does not instill declarative knowledge; rather, it enforces the specific structure, goals, and requirements of CN machine.

Failure Modes

We next examine the errors made by the foundational LLM and the SFT LLM when formulating CNs. As Figure 7 shows, the SFT LLM outperforms the foundational model across all failure modes. Specifically, fine-tuning effectively taught the SFT LLM to adhere to professional constraints by avoiding vague goals

(generic objectives lacking context) and compound statements (bundling multiple distinct CNs). Fine-tuning also instilled strict goals constraints, teaching the model to avoid solutions (technical implementations), processes (user actions rather than end states), and opinions (subjective stories or non-sequiturs). Our research partner notes that human analysts learn and transfer these same constraints; once mastered through training, they apply the constraints consistently across product and service categories.

Figure 7. Failure Modes for the Foundational LLM and the SFT LLM

6.6. Summary of Key Findings from Studies 1-3

Fine-tuning an LLM with proprietary professional judgments is efficient and effective. The CN machine automates the tedious, labor-intensive task of CN abstraction, enabling firms to surface comprehensive, high-leverage insights at scale. Because these performance gains are consistent across foundational models (including smaller architectures) and require modest training data, this capability can be democratized for innovators lacking the vast resources of larger firms.

Fine-tuning works by imparting task-specific inductive goals. Rather than relying on stored knowledge or product-specific data, the SFT LLM learns to transform source material into nuanced CNs that strictly adhere to professional goals and standards. The fine-tuned model does not retain general, declarative knowledge; instead it functions as a dedicated CN machine optimized solely for abstracting CNs from source sentences.

Finally, while the CN machine "forgets" its general-purpose functionality (Luo et al. 2025), this specialization is a feature, not a bug. The fine-tuned model excels at its intended task, while the

foundational LLM remains a superior choice for general reasoning.

7. Empirical Evaluation in Other Product Categories

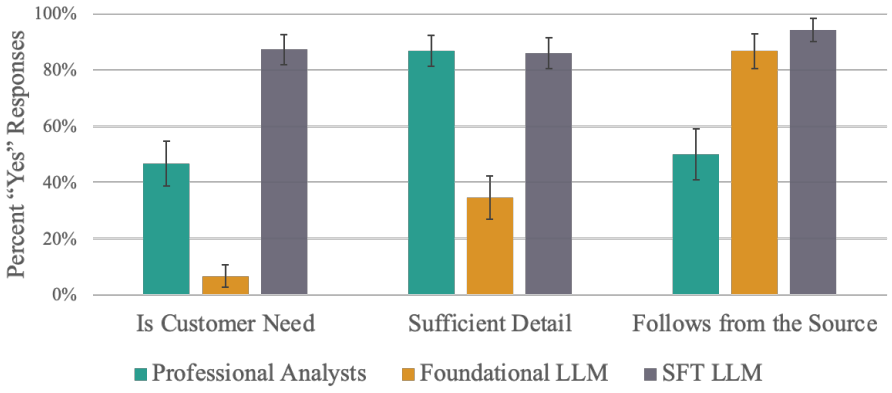
Studies 1-3 examined whether the CN machine meets professional standards in a controlled comparison. The applications in this section ask a different question: whether the CN machine lives up to those expectations under (1) different evaluative methodologies (Study 4) and (2) engagements with real clients, real data, and real decisions on the line (Study 5).

7.1. Study 4. Oral Care

We use data from TH 2019 to evaluate the SFT LLM's performance in the oral care category, following similar blind-test procedures as in Study 1a but using prior published data obtained via a different methodology. TH2019 began with 86 pre-identified CNs and tasked professional analysts with mapping them to a random sample of 8,000 Amazon review sentences. This CN-to-sentence task differs from the wood-stain sentence-to-CN task, thus affecting the professional-analyst benchmark. By analyzing these data, we examine whether the results of Studies 1-3 are methodology-dependent.

Figure 8 compares the performance of professional analysts, the foundational LLM, and the SFT LLM for oral care. The models remain the same, with no additional fine-tuning. Despite the differences between the wood-stain and oral-care studies, the SFT LLM yields a remarkably similar level of performance in absolute terms suggesting a stability across product domains and data structure.

Figure 8. Convergent test of Professional Analysts and LLMs in Oral Care



Consistent with Study 1, the SFT LLM is at least as effective as professional analysts. It performs substantially better on “Is Customer Need” ($z = 7.83, p < 0.01, h = 0.98$) and “Follows from the Source” ($z = 7.63, p < 0.01, h = 1.08$), and is not significantly different on "Sufficient Detail," with the point estimate favoring the SFT LLM ($z = 0.18, p < 0.86\ h = 0.02$). The professional analyst benchmark performs less well in oral care than wood stains, likely due to differences in how the benchmark was created (CN-to-sentence vs. sentence-to-CN for the benchmark).

Furthermore, the SFT LLM is substantially and significantly better than the foundational LLM on “Is Customer Need” ($z = 13.99, p < 0.01, h = 1.89$) and “Sufficient Detail” ($z = 9.08, p < 0.01, h = 1.11$), and it is slightly better on “Follows from the Source” ($z = 1.98, p = 0.49, h = 0.26$).

This convergent test in oral care is consistent with Study 1. The SFT LLM, but not the foundational LLM, is managerially and strategically relevant for automating the task of abstracting CNs.

7.2. Study 5. Product Development and Management Association (a Service Category)

Studies 1-4 are based on product categories and user-generated content. We now examine whether the SFT LLMs extend to a service category using experiential interviews (as in GH 1993). Our research partner was asked by the Product Development and Management Association (PDMA) to improve its membership services. The PDMA is the premier professional organization for product development (much like INFORMS is for management science), sponsoring national conferences, a journal, local chapters, and certification services. In cooperation with the PDMA, our research partner completed twenty 45-minute experiential interviews with current and potential PDMA members.

Professional analysts identified 512 CNs, while a Qwen-based SFT LLM – which does not experience fatigue – identified 1,153 CNs from the interview transcripts. Winnowing reduced the CNs to 97 analyst-identified CNs and 124 SFT-LLM identified CNs, which were further winnowed to a merged set of 143 tertiary CNs. Rather than a full three-level affinity diagram, our research partner structured these 143 CNs around 16 topics, such as “*Expertise*,” “*Certification*,” and “*Networking*.” The SFT-LLM-identified CNs spanned 87.5% of these topics, yielding slightly broader coverage than the professional analysts (81.3%).

The PDMA used the topics to guide the strategic discussions, while relying on the tertiary CNs to better understand context. The SFT LLM identified 70.6% of the 143 CNs, compared to 50.3% identified by professional analysts. As judged by our research partner, the professional-analyst-only CNs had little impact on the final managerial recommendations, and there was nothing systematic missed by the CN machine. Conversely, both the PDMA and our research partner judged the SFT LLM CNs to be better at identifying niche CNs – an important characteristic given the heterogeneity in the PDMA's customers.

For completeness, we ran the same blind professional evaluation used in Studies 1 and 4, asking professional analysts to judge whether each statement is a CN ("Is Customer Need") and whether it captures sufficient detail ("Sufficient Detail"). Professional analysts and the fine-tuned models all performed well and at similar levels on both questions (Online Appendix I). One finding in Study 5 stands out: while detailed few-shot prompting was not sufficient to match fine-tuning on "Is Customer Need" in the wood-stain category, it performed comparably to fine-tuning with the PDMA. We read this not as evidence against fine-tuning but as evidence for the concern that motivates it. The adequacy of prompting varies across categories in ways that are difficult to anticipate, whereas fine-tuning delivers consistent performance in every category we studied. At the current state-of-the-art, a practitioner cannot know in advance whether prompting is sufficient for a given category and data source. For example, it is possible that few-shot prompting succeeds here because the boundaries between CNs and solutions are more blurred in services than in physical products, or because experiential interviews provide greater context than UGC. Identifying the specific conditions under which prompting can substitute for fine-tuning is a promising direction for future work.

Managerial Implications

The final CNs provided fundamental insights that changed the way the PDMA operated. For example, the study revealed that veteran PDMA members knew one another and spoke in acronyms causing newcomers to feel like outsiders. Potential PDMA members expressed CNs for understandable communication and for rapid integration. Other CNs focused on personal gain: members wanted ideas and concepts to help them get promoted, get more respect in their fields, and improve their performance

as product-development professionals. The PDMA knew that community among potential members was important, but these findings suggested a different framing. The PDMA responded with new formats for formal events, methods to enhance employment opportunities, new member integration, a broad shift in the roles of members, and a way to facilitate inter-member communication. Many of these changes were introduced successfully at their premier annual meeting.

7.3. Other Applications

We briefly report four additional professional applications to the extent that we have received permission. (These are a subset of professional applications completed to date.)

*Food brand***.** This nationally advertised food brand effectively positioned itself around excellent taste, freshness, presentation, and valuable nutrients. The SFT LLM identified at least one set of new CNs – the customer desired the health value of the nutrients, not the nutrients *per se*. The product could be repositioned to improve a customer's hair, nails, joints, or other health and cosmetic outcomes.

Top-ranked specialized hospital. While high-quality care is a "must have,", identifying the family's CNs throughout the journey from specialist to specialist pointed to opportunities to improve the experience of all customers, not only the patient.

Composite siding. Based on 32 concept-testing interviews with homeowners, architects, and contractors (B2B), the SFT LLM abstracted 3,901 CNs, ultimately winnowed to 103 tertiary CNs. As in Studies 1, 4, and 5, the CNs identified by the SFT LLM were judged by our research partner to meet the same professional standards as those identified by analysts.

B2B paint and stain. While Studies 1-3 address B2C wood stain, this application demonstrated the feasibility for B2B applications based on professional forum data. New insights valued by the B2B paint-and-stain manufacturer were important to innovation in the category.

7.4. Qualitative Observations Based on Applications to Date

Formal evaluation of the effect of artificial intelligence on the role of professional analysts is an important research issue beyond the scope of this paper (e.g., Brynjolfsson, Chandar, and Chen 2025; Challapally et

al. 2025). Nonetheless, we provide qualitative observations to foster further research.

Across applications, we observe a consistent pattern in how the CN machine redivides expert work. The CN machine takes over the CN abstraction bottleneck, freeing analysts to elevate their focus to the judgment-heavy steps of innovation. Analysts still curate the source content (whether UGC or conducting interviews), strategically organize and prioritize the abstracted CNs, and facilitate cross-functional workshops to translate these "jobs to be done" into product roadmaps. These tasks require complex human judgment beyond current automation. Crucially, delegating abstraction to the machine breaks the traditional sample-size constraints of qualitative research. By processing entire corpora, without sacrificing diversity, the CN machine surfaces long-tail insights and niche segments that manual sampling routinely misses. VOC studies become more comprehensive, effective, and lower cost, putting professional-grade insights within reach of a broader set of firms and entrepreneurs. Ultimately, this transformation makes analysts more efficient and valuable to the firms, rendering their work more strategic, rewarding, and enjoyable.

8. Summary, Limitations, and Future Research

Across five interrelated studies, we provide convergent evidence that SFT LLMs perform at least as well as professional analysts in identifying customer needs. *A priori,* abstracting CNs is considered a challenging manual task that requires professional training in VOC and deep understanding of customer experience. The need for supervised fine-tuning was also not obvious. Given that LLMs excel at many unstructured, human-like tasks, it would not have been surprising if a foundational LLM performed well.

The fine-tuning process mimics how human professionals are trained, teaching the LLM the specific professional and content constraints of CNs. The SFT LLM does not rely on memorization. Rather, it becomes a specialized processor that transforms almost any input into the learned CN format, often at the expense of general-purpose reasoning. Because CN abstraction is a highly specialized task focused on short-form transformation, smaller-parameter models fine-tuned with relatively few examples appear sufficient to achieve professional-grade results. This puts the CN machine and high-quality insights

within reach for small firms and entrepreneurs with limited resources for consumer insights.

One question our design does not settle is how far future general-purpose models, combined with advanced prompting, could close the gap in CN quality. This is a question we find genuinely interesting for future research. Our studies establish that small models with modest fine-tuning can abstract CNs at a professional level. However, as frontier models evolve, future systems might require less fine-tuning, utilize a different form of fine-tuning, or reach professional quality through unanticipated mechanisms. We welcome such developments; our goal is reliable, accessible CN abstraction rather than any particular route to it. The CN machine is already transforming industry practice, and future AI advancements will further amplify the effectiveness and efficiency of consumer insights

8.1. Challenges Yet to be Met

The same mechanism that makes the CN machine reliable also defines its limits. Because fine-tuning installs a transformation skill rather than complex reasoning, tasks that require judgment rather than transformation remain human work for now.

While SFT LLMs master the abstraction of CNs, there are many steps in the VOC that remain human-centric. Winnowing and affinitization of (automatically) abstracted CNs is faster and easier than processing raw source sentences, but full automation remains an open question. We anticipate that solving abstraction will serve as a stepping stone toward automating downstream tasks.

Automating the prioritization of CNs is another open question. Researchers have attempted to use frequency, star-labeled ratings, and sentiment as proxies for importance, but such indicators are at best weakly correlated with CN importance and at worst counterproductive (GH 1993, TH 2019). Consequently, current practice relies on customer surveys. While the nascent field of "LLMs-as-respondents" holds promise, its findings are currently controversial and contradictory. For now, reliable prioritization by machine learning remains elusive.

References

Arora N, Chakraborty I, Nishimura Y (2025) AI-Human Hybrids for Marketing Research: Leveraging LLMs as Collaborators. *Journal of Marketing* 89(2): 43-80.

Autor DH (2014) Skills, education, and the rise of earnings inequality among the "other 99 percent". *Science* 344(6186): 843-851.

Baird A, Maruping LM (2021) The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS quarterly* 45(1): 315-341.

Brand J, Israeli A, Ngwe D (2023) Using LLMs for market research. *Available at SSRN* 23-062.

Brown SL, Eisenhardt KM (1995) Product development: Past research, present findings, and future directions. *The Academy of Management Review* 20(2):343-378.

Brown TB (2020) Language models are few-shot learners. *Advances in neural information processing systems 33: 1877-1901*.

Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. *Science* 358(6370): 1530-1534.

Brynjolfsson E, Chandar B, Chen R (2025) Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence, *Digital Economy*, Stanford University, Palo Alto, CA.

Bucher MJJ, Martini M (2024) Fine-tuned 'small' LLMs (still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660*.

Burchill G, Brodie CH (1997) *Voices into Choices: Acting on the Voice of the Customer*, Oriel Inc.

Arnould E, Cayla J, Beers R (2014) Stories that deliver business insights. *MIT Sloan Management Review* 55(2):54-62.

Challapally A, Pease C, Raskar R, Chari P (2025) The GenAI Divide: State of AI in Business 2025. *MIT Nanda*, Cambridge, MA.

Chen L, Zaharia M, Zou J (2024) How is ChatGPT's behavior changing over time? *Harvard Data Science Review* 6(2).

Chiang WL, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L et al. (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2, no. 3: 6.

Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y et al. (2024) Scaling instruction-fine-tuned language models. *Journal of Machine Learning Research* 25(70): 1-53.

Deroy A, Ghosh K, Ghosh S (2024) Applicability of Large Language Models and Generative Models for Legal Case Judgment Summarization. *Artificial Intelligence and Law*.

Dong G, Yuan H, Lu K, Li C, Xue M, Liu D, Wang W, Yuan Z, Zhou C, Zhou J (2023) How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Dong S (2024) Leveraging LLMs for Unstructured Direct Elicitation of Decision Rules. *Customer Needs and*

Solutions 11:1.

Gao L, Madaan A, Zhou S, Alon U, Liu P, Yang Y, Callan J, Neubig G (2023) Pal: Program-aided language models. *International Conference on Machine Learning*, pp. 10764-10799.

Gao Y, Lee D, Burtch G, Fazelpour S (2024) Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina. *arXiv preprint arXiv:2410.19599*.

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *MIS quarterly* 37(2): 337-355.

Griffin A (2004) Obtaining Customer Needs for Product Development. *The PDMA Handbook of New Product Development, 2E* (John Wiley & Sons, Inc.): 211-227.

Griffin A, Hauser JR (1993) The voice of the customer. *Marketing Science* 12(1):1-27.

Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, Li W et al. (2024) A survey on llm-as-a-judge. *The Innovation*.

Hauser JR (1993) How Puritan Bennett Used the House of Quality *Sloan Management Review*, 34, 3, (Spring), 61-70.

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *MIS quarterly* 28(1): 75-106.

Horton JJ (2023), Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research*, No. w31122

Ibrahim M (2025) Fine-Tuning LLaMa 2 for Text Summarization. *Weights and Biases*, February 22.

Ji J (2024) Demystify ChatGPT: Anthropomorphism around generative AI. *AI in Education, Culture, Finance, and War* 2:1.

Katz DM, Bommarito MJ, Gao S, Arredondo P (2024) Gpt-4 Passes the Bar Exam. *Philosophical Transactions of the Royal Society A* 382(2270): 20230254.

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS quarterly* 45(3): 1501-1526.

Lee D, Cheng Z, Mao C, Manzoor E (2025) Guided diverse concept miner (GDCM): Uncovering relevant constructs for managerial insights from text. *Information Systems Research* 36(1): 370-393.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5): 881-894.

Li P, Castelo N, Katona Z, Sarvary M (2024) Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science* 43(2): 254-266.

Lo CK (2023) What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13:4: 410

Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P (2022) Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*: 8086–8098.

Luo Y, Yang Z, Meng F, Li Y, Zhou J, Zhang Y (2025) An Empirical Study of Catastrophic Forgetting in

Large Language Models During Continual Fine-Tuning. a*rXiv:2308.08747v5*.

Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, Zettlemoyer L (2022) Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Mitchell JC (2016) What is a customer need? *Pragmatic Marketing*.

Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, and Pranav Rajpurkar (2023), "Foundation models for generalist medical artificial intelligence." *Nature* 616, no. 7956: 259-265.

Mullappilly SS, Shaker A, Thawakar O, Cholakkal H, Anwer RM, Khan S, Khan FS (2023) Arabic Mini-ClimateGPT: A Climate Change and Sustainability Tailored Arabic LLM. *arXiv preprint arXiv:2312.09366*.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine Your Own Business: Market Structure Surveillance Through Text Mining. *Marketing Science*, 31 (3): 521-543.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C et al. (2022) Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022): 27730-27744.

Panickssery A, Bowman S, Feng S (2024) LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024): 68772-68802.

Polanyi M (2009) The tacit dimension. *Knowledge in Organisations*, pp. 135-146. Routledge.

Qiu L, Singh PV, Srinivasan K (2023) How Much Should We Trust LLM Results for Marketing Research? *Available at SSRN 4526072*.

Rawte V, Sheth A, Das A (2023) A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Schick T, Dwivedi-Yu. J, Dessì R, Raileanu R, Lomeli M, Hambro E, Zettlemoyer L, Cancedda N, Scialom T (2024) Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36.

Schweidel DA, Moe WW (2014) Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research* 51(August):387-402.

Selinger E (2024) How to stop believable bots from duping us all. *Boston Globe* K1 November 24.

Shen J, Tenenholtz N, Hall JB, Alvarez-Melis D, Fusi N (2024) Tag-LLM: Repurposing General-Purpose LLMs for Specialized Domains. *arXiv preprint arXiv:2402.05140*.

Shi, Lin, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi (2025), "Judging the judges: A systematic study of position bias in llm-as-a-judge." *In Proceedings of the 14th International Joint Conference on Natural Language Processing*, pp. 292-314.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. *Nature medicine* 29(8): 1930-1940.

Timoshenko A, Hauser JR (2019) Identifying Customer Needs from User-Generated Content. *Marketing Science*, 38:1, 1–20.

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, et al. (2023) Llama 2: Open

foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Treutlein J, Choi D, Betley J, Marks S, Anil C, Grosse R, Evans O (2024) Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data. *38th Conference on Neural Information Processing.*

Van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: An ethnography of developing AI for hiring. *MIS quarterly* 45(3): 1557-1580.

Van Der Lee C, Gatt A, Van Miltenburg E, Wubben S, Krahmer E (2019) Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th international conference on natural language generation*, pp. 355-368.

Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A et al. (2024) Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *Nature Medicine* 30 (April):1134-1142

Yang Y, Uy MCS, Huang A (2020) FinBERT: A Pretrained Language Model for Financial Communications. *arXiv:2006.08097v2*

Zaltman G (1997) Rethinking Market Research: Putting People Back In. *Journal of Marketing Research*, 34 (4), 424–37.

Zhang J, Huang Y, Liu S, Gao Y, Hu X (2025) Do BERT-Like Bidirectional Models Still Perform Better on Text Classification in the Era of LLMs? *arXiv preprint arXiv:2505.18215*.

Zhang Z, Zheng C, Tang D, Sun K, Ma Y, Bu Y, Zhou X, Zhao L (2023) Balancing specialized and general skills in LLMs: The impact of modern tuning and data strategy. *arXiv preprint arXiv:2310.04945*.

Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z et al. (2024) Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36.

Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023) Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Transforming the Voice of the Customer by Automating Customer Need Abstraction

Online Appendices

Appendix A. Ten Product Categories Used for Model Training

Activewear	Housing and Apartments	Sleep Aids	Women's Underwear
Glucose Monitoring	Lawncare Equipment	Snow Removal Equipment	
Hearing Aids	Men's Shaving	Telehealth	

Appendix B. Examples of Training Data for Fine-tuning an LLM

Question	Answer
<GPT-VOC> <Product Category = "**Activewear**"> Just really curious why Oxford Gray on this is a different color than the Oxford Gray on the powerblend sweats.	Confident that colors will be consistent across products
<GPT-VOC> <Product Category = "**Recreational Vehicles**"> I tested it and it worked really well.	[]

Appendix C. Vicuna Fine-tuning and Inference

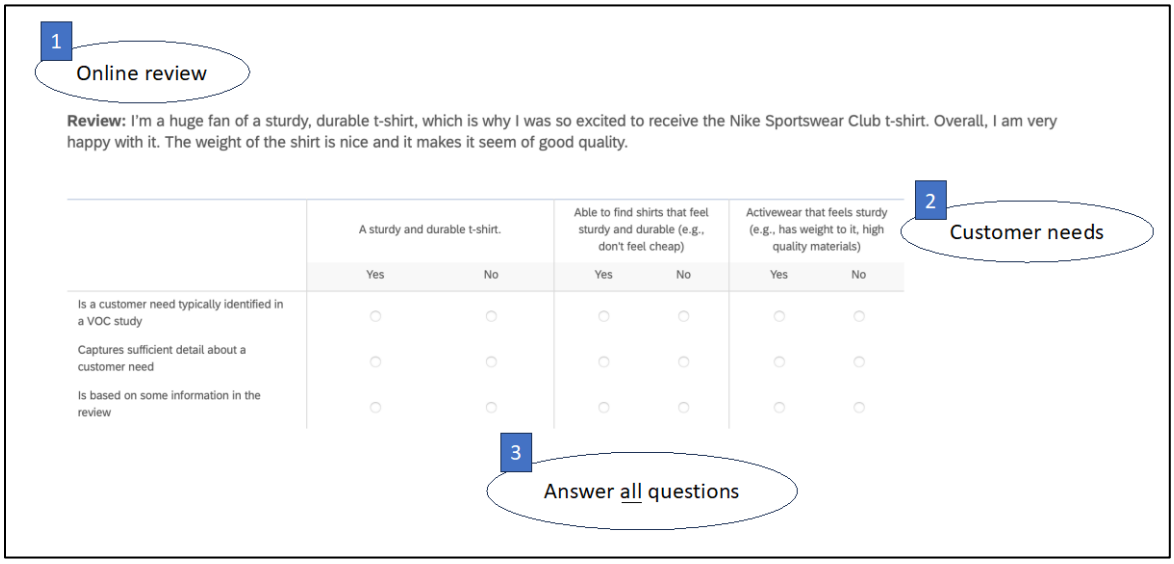
We fine-tuned Vicuna 13B using 4x Nvidia A100 PCIe on Lambda GPU Cloud. The system provides 40 GB VRAM, 120 vCPUs, 800 GB RAM, and 1 TB SSD storage. The fine-tuning process used bf16 precision without quantization, running for 6 epochs with a per-device batch size of 2 for training and 8 for evaluation. It employed a gradient accumulation of 4, a learning rate of 2e-5, cosine scheduling, and a maximum sequence length of 1024. Fine-tuning took approximately 8 hours. Inference, performed on 1x Nvidia A100 PCIe, takes about 0.4–0.5 seconds per prompt, or 400–500 seconds per 1,000 prompts.

Appendix D. Additional Examples of Customer Needs for Wood Stain Products

In Panel A, the professional analyst captures a CN perfectly, which demonstrates a deep understanding of the "job to be done." The SFT LLM abstracts a different CN. This CN is real, but it does not answer the question: Why do customers want to sand after the finish? The Base LLM reformulates generic concerns, instead of articulating a CN.

In Panel B, the professional formulation is a meaningful CN, but the idea differs from the original review. The Base LLM abstraction is a statement that does not look like a professional CN and misrepresents the information from the review. The SFT LLM yields the desired result.

Appendix E. Detailed Instructions in Blind Studies



Q1. Is a customer need typically identified in a VOC study.

Please indicate whether the statement qualifies as a customer need identified in a typical VOC study. Customer needs capture conceptual benefits that customers want to obtain from a product, which is

different from customer-provided technical specifications and desired solutions.

General Comment: For Q1, evaluate only if the statement is a customer need, regardless of whether the statement is detailed enough, which will be judged in Q2. This question also does not evaluate whether the statement came from the review, which will be judged in Q3.

Q2. Captures sufficient detail about a customer need.

Please evaluate whether or not the statement is actionable and not too general. For example, "good communication" might be too general. "Can stay informed of the technician's status (e.g. when they will arrive)" captures sufficient detail.

Q3. Is based on some information in the review.

Please evaluate whether or not the statement is based on information in the review. In particular, is it reasonable that a VOC study would abstract this customer need from the review?

Appendix F. Results for Study 1 Disaggregated by the Type of Source Content

	Verbatim			**Informative**			**Uninformative**		
	Prof. Analyst	Found. LLM	SFT LLM	Prof. Analyst	Found. LLM	SFT LLM	Prof. Analyst	Found. LLM	SFT LLM
Is Customer Need	0.81 (0.04)	0.33 (0.05)	0.92 (0.03)	0.80 (0.07)	0.20 (0.07)	0.90 (0.06)	0.93 (0.05)	0.17 (0.07)	0.93 (0.05)
Sufficient Detail	0.94 (0.02)	0.73 (0.05)	0.89 (0.03)	0.83 (0.07)	0.73 (0.08)	0.97 (0.03)	0.87 (0.06)	0.73 (0.08)	0.83 (0.07)
Follows from the Source	0.91 (0.03)	0.90 (0.03)	0.98 (0.02)	0.03 (0.03)	0.93 (0.03)	0.93 (0.05)	0.00 (0.00)	0.53 (0.09)	0.87 (0.06)

We highlight three observations. First, performance on "Follows from the Source," the for professional-analyst baseline drops close to zero for informative and uninformative reviews. This serves as an attention check in our survey design because we randomly selected professional-analyst-abstracted CNs for these reviews from the original VOC study. This does not affect Figure 3 in the main text because Figure 3 is based on the verbatim CNs only.

Second, it is not surprising that CNs identified by professional analysts from uninformative reviews were judged to be CNs with sufficient detail. Our design substituted other professionally identified CNs for this experimental cell. for, so as expected, evaluators agreed that the analyst-abstracted CNs are indeed CNs.

Third, we observe that the performance of the foundational LLM on the "Follow from the Source" dimension is lower for uninformative reviews than for informative reviews and verbatims. This might be hallucination. The original professional analysts found these reviews uninformative in the professional

VOC study, so this result is expected. However, the SFT LLM does remarkably well reinforcing its ability to perform at least as well, perhaps better than, professional analysts.

Online Appendix G. Transformer-based Classifier

To maximize accuracy, we trained separate classifiers for each of the three evaluation questions in Study 1 (wood stain). The classifiers for "Is Customer Need" and "Sufficient Detail" are based on the output CNs in isolation. The "Follows from the Source" classifier assesses sentence-CN pairs. We benchmarked two foundational LLMs (Claude Sonnet 3.7 and ChatGPT 4o) and three pretrained transformer models (ModernBERT, RoBERTa, and BERT), and found that RoBERTa optimized with varied asymmetric costs and focal loss, yielded the best cross-validation performance across all questions. The superior performance of fine-tuned classifiers relative to LLMs when sufficient labeled data are available is consistent with existing literature (Bucher and Martini 2024; Zhang et al. 2025).

Figure G1 demonstrates that the classifiers effectively replicate professional evaluators. (Recall that the error bars are confidence intervals, not standard errors.) Using an 80/20 cross-validation split of professional-analyst and foundational LLM data from Study 1, the out-of-sample predictions show no statistically significant differences from human evaluations across the nine metrics (using Bonferroni corrections). Notably, the classifiers accurately predict performance for the SFT LLM outputs, despite all SFT LLM examples being excluded from the classifiers' training. These results confirm that the classifiers are reliable proxies for extracting qualitative insights.

Figure G1. Evaluating Transformer-based Classifiers for Questions in Study 1

Appendix H. Study 2 Results for "Sufficient Detail" and "Follows from the Source"

Figure H1. Predicted Performance with Different Amounts of Fine-Tuning Data

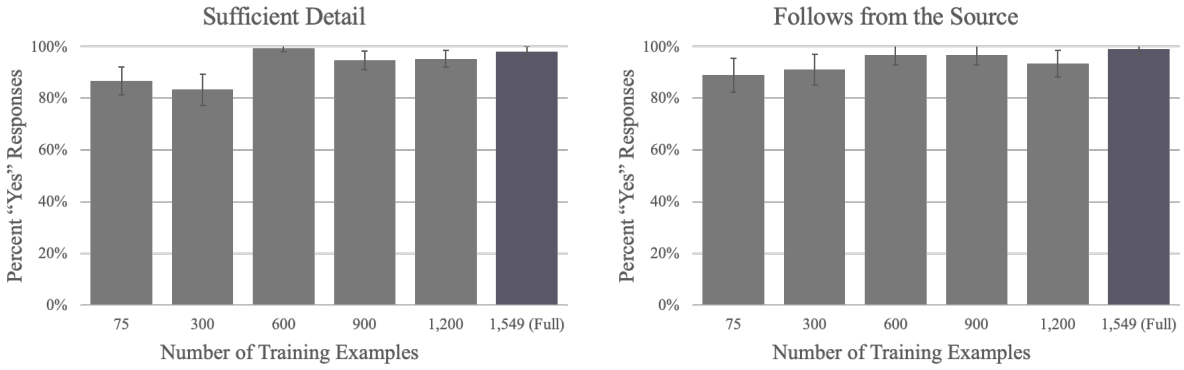
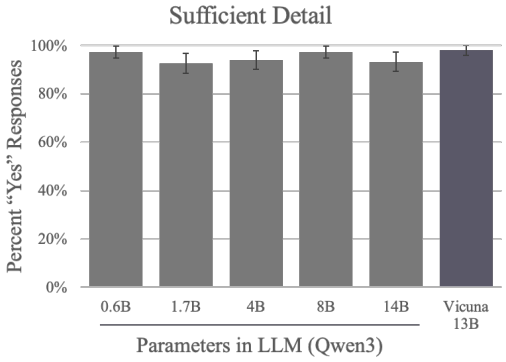


Figure H2. Predicted Performance with 600 Training CNs from Different Number of Categories

Figure H3. Predicted Performance of Different Foundational LLM Architectures



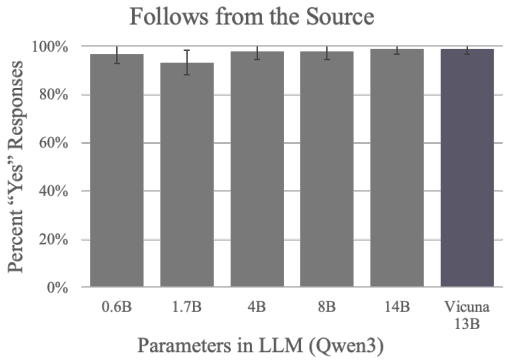


Figure H4. Predicted Performance of Alternative Prompting Strategies

Appendix I. Results for Blind Evaluations with the PDMA Service Category