Can Large Language Models Extract Customer Needs as well as Professional Analysts?

by Artem Timoshenko, Chengfeng Mao, and John R. Hauser

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, E62-510, 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.

Acknowledgements

We thank Carmel Dibner, Kristyn Corrigan, John Mitchell, Rachael Settipani, and Maggie Hamilton at Applied Marketing Science, Inc. for insightful discussions and research collaboration. 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, and research seminars at University of Texas at Dallas, Colorado University at Boulder, and the University of Florida.

Disclosures The authors have no actual or perceived conflicts of interest.

The paper uses proprietary data about customer needs in different industries. We cannot disclose these data. We propose option (a) from JMR's alternative data disclosure plan: Disguise the data to protect sensitive information but allow replication of the main results. We will replace the exact formulations of customer needs with unique identifiers. This would allow replication of all statistics reported in the empirical evaluation.

Can Large Language Models Extract Customer Needs as well as Professional Analysts?

Abstract

Identifying customer needs (CNs) is important for product management, product development, and marketing. Applications rely on professional analysts interpreting textual data (interview transcripts, online reviews, etc.) to understand the nuances of customer experience and concisely formulate "jobs to be done." The task is cognitively complex and time-consuming. Current practice facilitates the process with keyword search and machine learning, but relies on human judgment to formulate CNs. The paper examines whether Large Language Models (LLMs) can automatically extract CNs. Because evaluating CNs requires professional judgment, the authors partnered with a marketing consulting firm to conduct a blind study of CNs extracted by: (1) foundational LLM with prompt engineering only (Base LLM), (2) LLM fine-tuned with professionally-identified CNs (SFT LLM), and (3) professional analysts. The SFT LLM performs as well or better than professional analysts when extracting CNs. The extracted CNs are well-formulated, sufficiently specific to identify opportunities, and justified by source content (no hallucinations). The SFT LLM is efficient and provides more complete coverage of CNs. The Base LLM was not sufficiently accurate or specific. Firms can rely on SFT LLMs to reduce manual effort, enhance the precision of CN articulation, and provide improved insight for innovation and marketing strategy.

Keywords: Voice of the Customer, Customer Needs, Marketing Research, Product Development, Innovation, Machine Learning, Generative AI, Large Language Models

Introduction

Understanding customer needs (CNs) is essential for effective product development, product management, and marketing strategy. For example, a snowplow company recognized that when customers plow the sidewalks, they often need to turn from one narrow perpendicular sidewalk to another. This insight motivated them to develop a zero-turn snowplow brand that immediately solved an important CN. The movie theater business was revolutionized with stadium seating to fulfill the CNs of a clean & unobstructed view, room to rest arms, spacious well-cushioned seats, can stretch legs, easy to get in and out, rest my head, and storage for refreshments. The identification of patient and physician CNs (such as easy to interpret diagnostic information, convenient-sized output, and easy to hold) 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 a sufficiently abstract level to indicate what the product (or service) needs to accomplish, rather than which specific product attribute fulfills a CN: clean, unobstructed view in a movie theatre rather than a 26" wide, swivel chair with cupholders positioned three feet to 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 then 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 task. Current industry practice involves manually analyzing qualitative data such as interview transcripts, online reviews, and call center data. Because customers rarely articulate CNs explicitly, professional analysts must interpret customer statements to distill insights into concise, clear statements that capture the full nuance of CNs. Interpretation and formulation require both skill and patience.

Analysts are trained to develop a deep understanding of the customer's experience to recognize and articulate CNs accurately. However, the manual approach does not scale well and slows the time-to-market.

The abstract and context-dependent nature of CNs means that automating CN extraction is challenging. Current automation approaches involve keyword searches and the machine-learning identification of informative and diverse content to augment human judgment (Timoshenko and Hauser 2019). These methods help firms to screen large qualitative data, but the critical final step of precisely formulating CNs continues to rely on human expertise.

Large language models (LLMs) show promise. LLMs formulate coherent sentences and have achieved success in abstractive summarization tasks (e.g., Arora, Chakraborty, and Nishimura 2025). Unlike human professionals who find the CN identification task tedious, an LLM does not lose focus due to fatigue, can scale to larger source material, and speed time-to-market. By automating CN identification, a well-performing LLM would process more customer information to provide better insights and would free human professionals to focus on generating creative solutions for product development and marketing.

Whether LLMs can formulate CNs for professional applications remains a question. Industry adoption requires precision in capturing the nuanced CNs (not too general, not too specific). CNs must also be formulated correctly to effectively communicate customer insights to product managers. Simple rephrasing is inefficient and prompt engineering does not appear up to the task (see related research by Gao et al. 2024). On the other hand, supervised fine-tuning (SFT), in which the LLM's parameters are finetuned with manually curated training examples, could help to address these challenges (Dong et al. 2023, Lewis et al. 2020). But SFT potentially requires a large number of professional examples of extracted CNs. Such data are often

proprietary and the task is not well-suited to an online workforce marketplace such as Amazon Mechanical Turk, Prolific, or Lucid (Cint).

Evaluation is a challenge. We must evaluate any extracted CNs with respect to the deep insights rather than superficial natural language. For example, the wood-stain-product tertiary 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," "can achieve a glossy or flat finish, depending on my preference" and other phrases in our data. Whether or not these phrases represent CNs, provide the necessary specificity, and are true to the source material remains a professional (human) judgment task. CNs capture the underlying benefits of the product or service, and they can be confused with target values, solutions, and customer opinions. Evaluation requires trained professionals familiar with the CN elicitation task for applications. Ground truth is hard to obtain, in part because LLMs are easy to anthropomorphize and provide CNs that sound right to the untrained ear (Ji 2024, Selinger 2024).

This paper investigates whether an SFT LLM can extract customer needs from online reviews and interview transcripts. The evaluations are conducted in a blind study with a professional marketing consulting firm. Benchmarks include professional analysts and a foundational LLM with prompt engineering only (Base LLM).

Related Literature

Identifying Customer Needs (CNs)

CNs are the basis of the voice-of-the-customer (VOC, Griffin and Hauser 1993; hereafter, GH

1993). Since that paper was published, there have been hundreds of academic and industry articles on improved methods for qualitative interviews, ethnographic methods, metaphor elicitation, and interpretation (e.g., Brown and Eisenhardt 1995, Burchill and Brody 1997, Gupta 2020, Mitchell 2016, Zaltman 1997, Cayla, Beers, and Arnould 2014). All proposed methods require human judgment to interpret customer interviews.

More recently, firms recognized that user-generated content (UGC; e.g., online reviews, blogs, and forums) augments customer interviews, which requires new methods to scale the customer need analysis. Initially, research focused on the word counts, word co-occurrence, and topic models to identify "bags of words," but "bags of words" cannot describe nuanced CNs (Lee and Bradlow 2011, Netzer et al. 2012, Schweidel and Moe 2014, Büschken and Allenby 2016). To capture CNs, Timoshenko and Hauser (2019; hereafter, TH 2019) use convolutional neural networks to identify diverse informative sentences that can be reviewed by professional analysts to extract CNs. Our paper investigates whether LLMs can automate this last step and formulate CNs from qualitative data as well as professional analysts.

Large Language Models

Large Language Models (LLMs) use deep and parallel neural network layers and are pretrained on vast amounts of text data to understand, generate, and respond to natural human language. Current LLMs, such as GPT-4o, Claude, and LLaMA 2, are based on the transformer architecture (Achiam et al. 2023, Touvron et al. 2023). The self-attention modules in transformers handle sequential data at scale, allowing for parallel processing and capturing long-range dependencies in text (Vaswani et al. 2017). The LLMs are typically trained using a combination of self-supervised learning and reinforcement learning from human feedback (Christiano et al. 2017).

LLMs have demonstrated remarkable capabilities across domains, such as education (Kasneci et al. 2023, Lo 2023), healthcare (Moor et al. 2023, Thirunavukarasu et al. 2023), coding (Gao et al. 2023), and law (Katz et al. 2024). The marketing science community recently started to explore applications of LLMs for marketing research. One prominent idea is that LLMs can serve as synthetic respondents (Horton 2023). For example, Arora, Chakraborty, and Nishimura (2025) use LLMs to create marketing personas that answer qualitative and quantitative questions. Brand, Israeli, and Ngwe (2023) apply LLMs to obtain willingness-to-pay estimates and reproduce conjoint studies. Qiu, Singh, and Srinivasan (2023) evaluate LLMs for eliciting consumer risk preferences. Li et. al (2024) explore LLMs for automated perceptual mapping. Dong (2024) uses LLMs to successfully replicate the customer decision rules identified by human judges in unstructured direct elicitation (Ding et al. 2011).

LLMs do well on many tasks, but not all tasks. Gao et al. (2024, p. 2) suggest that many LLMs "differ markedly from that of human participants" and "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. Prompt engineering is often effective (Brown 2020), but it does not always work and it is a challenge to consistently generate high-quality prompts (Min et al. 2022). Lu et al. (2022) demonstrate that examples provided in prompt engineering are not always effective; differing orders of prompts can either result in excellent or random-guess performance. For human-oriented decisions, Gao et al. (2024) demonstrate that simple queries, prompt engineering, and providing external documents as references (retrieval automated generation, RAG) differ in distribution from human respondents. Arora et al. (2025) suggest that, for quantitative studies, LLM-based answers are directionally reasonable but often underestimate

respondent heterogeneity. In our study, along with the prompt-engineering approach, we examine an LLM that is fine-tuned using a database of professionally-identified CNs. The SFT approach allows us to standardize the prompt structure for our specific problem and train the SFT LLM model to formulate CNs using established industry-standards.

Industry Practice

Before we evaluate whether an SFT LLM and/or a Base LLM can match the performance of professional analysts, we need to understand the professional-analysts' task.

Formulating CNs is demanding. The challenge lies in understanding the deeper motivations that drive customer behavior and capturing these motivations in a concise and efficient form. Industry professionals often differentiate between CNs, solutions, targets, and opinions. For example, a customer might express dissatisfaction with battery life in cellphones (an opinion), but the underlying CN might be the desire for longer, uninterrupted use while traveling. In the academic literature, solutions and targets are often framed as product attributes, while opinions reflect customer sentiments. Understanding deep CNs before focusing on specific solutions provides insights beyond the current market offerings. Before the zero-turn snowplow was launched, there was no snowplow that could move from one perpendicular sidewalk to another. The ability to do so was not a defined attribute. Before stadium seating was introduced to movie theaters, attributes such as drink holders and elevation were not part of the conversation.

To identify CNs, firms use experiential interviews, metaphor elicitation, ethnography, focus groups, call center logs, or user-generated content to create a corpus of sentences. Analysts highlight relevant sentences and phrases in the source material and rephrase the customers' words as CNs, keeping the verbiage and the customer's intentions as close as feasible to that

articulated by the customer.

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 extracted a CN "*Able to know the right amount of time to spend on each step of my oral care routine.*"

CNs tend to be nuanced, yet not too specific to forestall creativity.

Example 2. A professional study for oral care products 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*."

Extracting CNs is cognitively demanding requiring analysts to translate raw customer language into precise, actionable CNs. A wood-stain customer might say: "*Always great to use. I do wish stores would stock the amber color in the gallon size.*" An analyst would formulate a customer need as "*Able to find the size and color combinations that I want in store."* CNs balance specificity and generality. 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.

Humans are fallible. GH 1993 report that each analyst was able to identify 54% of the CNs (range 45-68%) that were ultimately identified. With more applications, professional analysts have become more skilled, but still not perfect. New analysts receive training materials that define CNs, contrast CNs with (existing) solutions, and provide standards to extract 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 an SFT LLM.

Depending on the product (service) category and the source material, analysts might identify hundreds or even thousands of potential CNs. Not surprisingly, there is redundancy. Using keyword matching and experience-based judgment, analysts “winnow” the CNs to a smaller, less redundant set. To focus product management, product development, or marketing on creative solutions, analysts create a hierarchy of primary, secondary, and tertiary CNs (Burchill and Brody 1997, GH 1993). While creating the hierarchy (“affinitization”) the CNs are winnowed further. The winnowing and affinitization tasks require training and experience to channel customers as informed by the source material.

Large Language Models for Extracting Customer Needs

From a computer science perspective, formulating CNs requires abstractive summarization. Abstractive summarization involves generating new phrases that capture the core meaning of an input in a more conceptual and concise manner. For instance, an abstractive summary of a news article might condense complex details into "*World leaders discussed global strategies to mitigate climate change and reduce emissions*," even if those exact words don’t appear in the original text. In contrast, extractive summarization focuses on identifying and reproducing key phrases, such as when a search engine extracts snippets containing exact sentences from documents. LLMs are well-suited to the more-challenging task of abstractive summarization.

To examine whether LLMs can extract deep nuanced CNs, we develop prompts for a Base LLM and gather training data for an SFT LLM. We then use (blinded) professional analysts to

evaluate CNs produced by the Base LLM, the SFT LLM, and other professional analysts.

Foundational Model: Vicuna 13B

The foundational model in our analysis is Vicuna 13B (Base LLM). In our preliminary investigation, CNs formulated by Vicuna 13B were qualitatively similar to the ones formulated by the state-of-the-art publicly available models (including Chat GPT-4). We used Vicuna 13B because the smaller Vicuna 7B model did not perform as well, and there was little further improvement for the larger Vicuna 33B. Importantly, Vicuna 13B offers a license for academic use, which enables reproducibility. Because our research requires professional analysts to read and evaluate model outputs, an expensive and time-consuming process, we were required to commit to a model architecture early in our research.

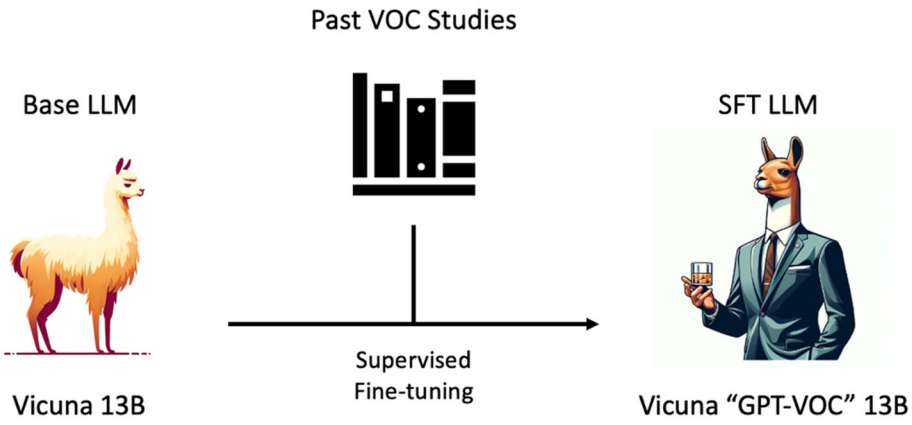
Vicuna is a general-purpose open-sourced LLM, developed by academic researchers from UC Berkeley, UCSD, and CMU (Chiang et al. 2023). Vicuna uses LLaMA 2 as a base model and finetunes it with 70,000 user-shared ChatGPT conversations (Touvron et al. 2023). Vicuna performed comparable to the current open-source 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).

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 analysis, all options performed similarly. Overall, the best prompt was: "*For a <Product Category>, identify a customer need from the user review. If no need is found, return []. Review: <Review Text>*"

Finetuning an LLM with Professional VOC Studies

Just as professional analysts gain expertise by recalibrating based on feedback, we trained an SFT LLM to extract CNs with feedback from professional voice-of-the-customer studies. See Figure 1.

Figure 1: Finetuning an LLM with CNs from Professional VOC Studies



The training data span ten different product categories, such as activewear, glucose monitors, and recreational vehicles. The data from each product category came from VOC studies by a firm with over 30 years of experience in market research. Each application provided valuable information to a client who either improved products and services or improved their marketing. Appendix A provides a complete list of product categories. For each product category, the firm reviewed interview transcripts and/or online reviews and identified CNs. Our data also include matched verbatims (such data are rare). We augment these data with "negative" examples that include sentences from interviews and online reviews that contain no information about customer needs (uninformative sentences).

Figure 2 illustrates two examples from our training data. The question-and-answer structure is similar to the instruction finetuning in natural language processing (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 at the beginning of each prompt to condition the model on identifying CNs for the category.[1] It is a common practice to use a special tag to condition the LLM with a new task, such as using tools (Schick et. al 2024) or predicting protein (Shen et. al 2024). After finetuning, we can use the standardized prompt as a shorthand to indicate to the SFT LLM that it is to extract CNs.

Figure 2: Examples of Training Data for Finetuning 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.	[]

During finetuning, the SFT LLM attempts to extract CNs and receives feedback on the "correct" output – either the CN or []. We use backpropagation to finetune the parameters of the model. See Dong et al. (2023) for details on finetuning. Finetuning was facilitated by the DeepSpeed 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 are run on a local workstation (desktop).

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%).

[1] We used "GPT-VOC" as a reference to "Generative Pretrained Transformer for the Voice-of-the-Customer." The special tag should be unique, but the tag is not required to be meaningful.

Additionally, the transcripts contained 11,975 uninformative sentences. We need "negative" examples to train the model to output "[]" for uninformative content, but we must choose the number of negative examples carefully to control the tradeoff between false negatives and hallucination. Too many negative examples lead to missing CNs (false negatives), while too few negative examples lead to CNs that do not follow from the verbatims (hallucination). We selected the number of randomly-chosen negative examples (47) by observing the model performance on the validation data. Figure 2 provides examples of both positive and negative samples.

After finetuning, the CNs identified by the SFT LLM appear to capture the CNs articulated in the reviews (no evidence of hallucination) and span a broad range of customer insights, including rarely-mentioned niche customer needs. We now test that proposition.

Empirical Evaluation of Professional Analysts vs. LLMs

Evaluating the veracity, relevance, and comparability among methods of rich verbal output is always a challenge. For CNs, the evaluation requires well-trained analysts to read the stated CNs and judge whether these statements adhere to professional standards (Griffin 2004, also the PDMA's Glossary for New Product Development). Although analyst training might vary firm-to-firm, our research partner uses extensive training and peer support to ensure that the definition of CNs is consistent among professional analyst judgments. Our research partner has conducted numerous VOC studies for consumer brands and business-to-business organizations, and is often called upon to train other firms in CN identification. To evaluate CNs, we recruited professional analysts from the same firm (our research partner) that provided the training data. The analysts who participated in our research were not involved in the initial for-client VOC studies used in

model training and evaluation. Our methodology assured the analysts were blind to whether the customer needs were formulated by other analysts, the Base LLM, or the SFT LLM. In total, the value of professional time donated by the firm and the professional analysts was substantial.

Our primary evaluation focuses on wood stain products, and we discuss additional applications with oral care products and a professional association.

Data Overview: Wood Stain Products

For the wood stain product category, after winnowing, the original VOC study identified 103 CNs. The analysts 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). The 1,000 number was selected in a typical cost-versus-quality tradeoff for a client-based VOC application. Following standard procedures, the analysts read the selected reviews and manually extracted the unique CNs. The firm shared with us the verbatims for every extracted CN. We applied the Base LLM and the SFT LLM to identify CNs from the same online reviews. Both LLMs are automated and scale well to all 14,341 review sentences.

To minimize information leakage, the wood stain category was not used in LLM finetuning. Although some information on wood stains might have been available during the foundational training of Vicuna 13B, professional VOC studies and CN formulations are trade secrets and unlikely to be available publicly.[2]

Illustrative Example of Wood Stain Customer Needs

Figure 3 illustrates the output of the LLMs for wood stain products. From the source material

[2] Data leakage might be an explanation if the Base LLM does well relative to human analysts and the SFT LLM. Our findings suggest this is not the case. Any leakage in foundational training for Vicuna 13B would only reinforce the key qualitative recommendations in this paper.

(online review) in Figure 3, professional analysts extracted the CN: "*Able to see what surface areas I have already covered*." This CN seems relevant when wood stain is applied to larger surface areas.

Figure 3: Illustrative Example Customer Needs Identified from Online Reviews

Without finetuning, the Base LLM extracted a CN "*Easy to see coverage*." While this statement correctly summarizes the topic of the online review, the statement lacks the specificity required to inspire innovation. The performance of the Base LLM in Figure 3 is typical; the Base LLM often paraphrases the customer review or focuses on solutions and opinions.

The SFT LLM extracted a CN "*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 contains similar clarifications in 34% of the examples. While the CN extracted by the SFT LLM is different from the CN extracted by a human analyst, the CN captures similar meaning and adheres to similar professional standards—a judgment we examine formally in our study.

Appendix C provides additional examples of online reviews and the corresponding CNs as

extracted by professional analysts, the Base LLM, and the SFT LLM.

Study 1. Are Statements Extracted by LLMs Well-Formulated Customer Needs?

Our first study evaluates whether CNs extracted by analysts and LLMs adhere to the industry's professional standards. Three professional analysts, not involved in the original wood stain application and with experience in VOC studies, evaluated CN statements on three dimensions. The wording of the questions to the analysts was based on extensive discussion with the firm's VOC experts. The questionnaire was pretested and revised so that each question was clear, understandable, and measured the target construct. Appendix D provides detailed instructions, including the user interface of the study design and clarifications about the evaluation dimensions. 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 ("Sufficiently Specific")

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

Each analyst evaluated 150 randomly-chosen sentences from online reviews. For each review sentence, the analyst was given the text of the online review and three CN statements ([other] professional analyst, Base LLM, and SFT LLM). We randomized the order of CNs for each review. Analysts were blind to the purpose of the study and posttests indicated that there were 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 review *x* CN combination.

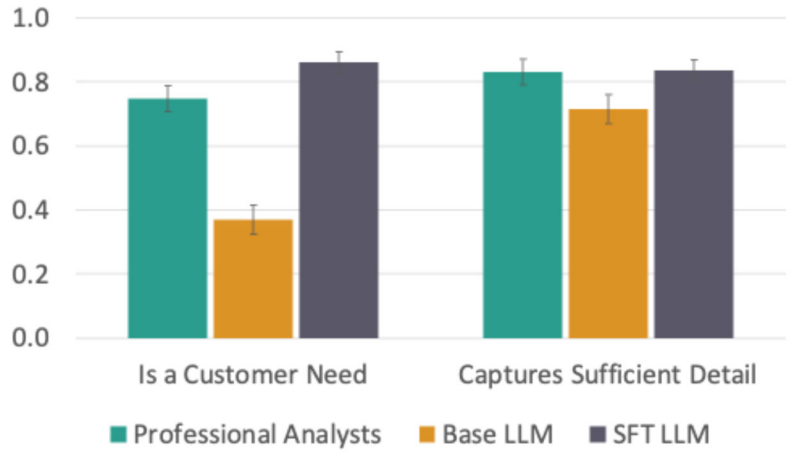
The sample of 150 review sentences included (1) 90 sentences indicated as *verbatims* leading to CNs in the voice-of-the-customer application, (2) 30 reviews indicated by the firm as *informative* but not used as verbatims, and (3) 30 reviews indicated by the firm as *uninformative*.

The firm considered the uninformative reviews in the original for-client VOC study and decided that they do not contain CNs. For *verbatims*, all three approaches identified CNs. For the other categories of reviews, the analysts in the original VOC study decided not to formulate CNs. To maintain blinding, every question contained three plausible CNs to avoid any inadvertent signals about how a CN was extracted. We achieved three plausible CNs by augmenting the Base-and-SFT-LLM-identified reviews with analyst-identified CNs randomly selected from the original VOC application.[3]

Figure 4 reports results for the first two questions, aggregated across the verbatim, informative, and uninformative reviews. (Appendix E reports the disaggregated results.) The professional-analyst baseline provides an important benchmark for the LLMs. Professional analysts extracted these customer needs in a for-client application. Despite extensive training, professional analysts are not perfect. In 1993, Griffin and Hauser reported imperfect (54%) identification of CNs from experiential interviews. In 2024, the evaluative analysts agreed with the original analysts about 80% of the time on "Is Customer Need" and "Sufficiently Specific." Although the task in Figure 4 is not identical to the Griffin-Hauser task, the 80% agreement suggests improvement in industry practice and reinforces our decision to rely on professional analysts rather than research assistants or an online workforce marketplace.

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

Figure 4: Comparison of Customer-Need Extraction by Professional Analysis and LLMs



The SFT LLM is particularly effective. In our study, the SFT LLM identifies CNs that are significantly more likely to be judged as CNs than those identified by the original (fallible) human analysts ($p < 0.01$). The CNs identified by the SFT LLM are as specific as the CNs identified by the original analysts ($p = 0.86$). This result is important for practice. Figure 4 suggests that an SFT LLM could automate the tedious task of extracting CNs from source material. Analysts are freed to focus on higher value-added tasks.

Figure 4 cautions that a Base LLM, even if provided examples and carefully prompted, does substantially and significantly worse than both human analysts and the SFT LLM when extracting CNs ($p < 0.01$ and $p < 0.01$). The specificity of the Base LLM is relatively better, but still significantly less than either professional analysts or the SFT LLM ($p < 0.01$). Figure 4 suggests that quality training data and supervised finetuning enable an LLM to be used to extract CNs. Without the training data and finetuning, the Base LLM does not do well.

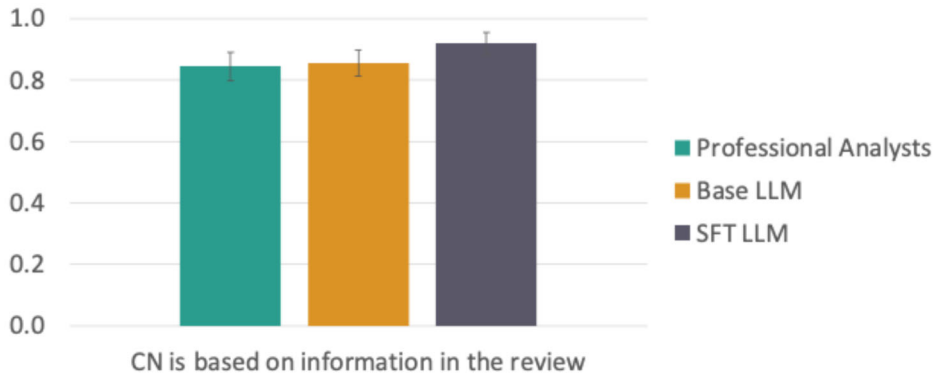
We next examine whether hallucinations are a problem for either of the LLMs (Rawte et al. 2023), or for human analysts completing the tedious task of reading through thousands of online reviews. To extract CNs, analysts and LLMs interpret information in the verbatims to understand

the customer's stated "job to be done." Customers do not state CNs directly. There is room for interpretation and that interpretation might inadvertently be based on other information in the entire corpus of interviews or reviews.

In Figure 5, we evaluate whether the extracted 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 extract this customer need from the review?"

Figure 5 focuses on the 90 *verbatims.* Professional analysts in the original study judged the v*erbatims* to provide CNs that are based on information in the reviews. Limiting Figure 5 to verbatims provides a fair comparison among professional analysts and the LLMs and, if anything, favors professional analysts. The appendix reports data from all other review sentences. The data are face valid and reinforce our interpretations.

Figure 5: CNs Identified by Analysts and LLMs Capture Information from the Reviews



The analyst-extracted CNs were judged to closely represent information from the verbatims for 84% of observations. The firm reports that this percentage is within an expected range for VOC applications. The LLMs outperform analysts slightly, with the SFT LLM performing best; this difference is significant for the SFT LLM ($p < 0.01$) but not for the Base LLM ($p =$

0.72).

Figure 5 is an important observation. VOC studies are designed to summarize CNs as expressed by customers. Figure 5 suggests that LLM hallucinations are not a practical problem for CNs. Figure 5 provides further evidence that SFT LLMs can free analysts to focus on other aspects of providing managerial advice.

Study 2. Can LLMs Capture Diverse Customer Needs?

Before CNs are used to improve strategies and tactics, CNs are grouped into an affinity diagram—a hierarchical structure of primary CNs, secondary CNs, and tertiary CNs (Burchill and Brody 1997; GH 1993). Redundancy among CNs is reduced and higher-level CNs (primary and secondary) are chosen to represent the customer's perspective and to ensure that the higher-level CNs provide sufficient breadth for managerial application. For example, for wood stain products, the CN "a*ble to control depth and finish of the stain and topcoat*" could belong to the secondary CN "*application looks even and consistent,*" from the primary CN "*appearance and finish*." To evaluate whether the LLMs capture a broad set of CNs, we require a professionally-developed affinity diagram based on a concatenation of the professional-analyst and SFT LLM CNs.

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 14,341 source verbatims. Analysts used the same procedure used for professional VOC applications 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 during the affinitization process, the analysts further winnowed the CNs to a final set of 117 CNs. The analysts were blind to the source of customer needs (original analysts or SFT LLM), but we preserved the mapping from the 257 (154 plus 103) CNs to the final 117 CNs.

Step 3. Mapping CNs Back to Verbatims: Our research partner reconstructed the mapping from a randomly-sampled set of 2,000 SFT-LLM-identified CN statements to the final 117 CNs. For each of the 2,000 SFT-LLM-identified statements (pre-winnowing), analysts mapped the 2,000 statements to the final 117 CNs. The mapping is many-to-many because, consistent with Figure 4, some LLM-formulated statements were judged not to be CNs. Many remaining CNs were eliminated as redundant or merged during the winnowing process.

The mapping required a major effort over several months, and it would have been cost-prohibitive to reconstruct a mapping for all CNs identified from the entire set of 14,341 source verbatims. In typical VOC applications, the mapping from final CNs to source verbatims is not maintained because the mapping is not considered valuable for business applications. The substantial cost of reconstructing the mapping is rarely justified by any corresponding benefits. Indeed, in over thirty years of VOC applications by our research partner, we know of only one instance in which the firm recreated the mapping—a litigation support application which required extensive documentation.

Strategic Value: Managerially Relevant Primary and Secondary Customer Needs

For wood stain products, the final affinity diagram includes 30 secondary CNs which were in turn grouped into eight primary CNs. The SFT-LLM-identified CNs had slightly more strategic breadth. The SFT LLM identified CNs from 100% of the secondary groups and 100% of the

primary CN groups. The professional analysts identified CNs from seven (87.5%) of the primary groups and 24 (80%) of the secondary groups. For example, the analysts in the original study missed the primary CN which describes *a desire for products that help make the maintenance of wood easier*.

Looking at the tertiary CNs, the 154 winnowed SFT-LLM-identified CNs account for 84.6% of the final affinitized tertiary CNs while the 103 winnowed analyst-identified CNs account for 48.7%. (The percentages add to more than 100% because of overlap.) These statistics must be interpreted cautiously. Typical cost-vs.-benefit considerations limited the original VOC study to a sample of the corpus, but the SFT-LLM scales well to the full corpus. Second, these percentages confound redundancy reduction (both stages of winnowing) with whether an after-winnowing-and-affinitization tertiary CN is equivalent to one of the CNs extracted from the source material. We now address that confound by returning from the winnowed CNs to the source verbatims.

How Many Online Reviews are Sufficient for a Voice-of-the-Customer Application?

TH 2019 suggest that online reviews are a rich source of CNs for oral care products. From 8,000 online review sentences (approximately 4,200 informative sentences), professional analysts identified 87% of the oral care CNs in the final affinity diagram. We construct a similar statistic to evaluate the ability of the SFT LLM to identify the CNs in the wood stain affinity diagram. For wood stain, the SFT LLM rather than analysts decides which sentences are informative and formulates CNs. It implicitly and automatically detects uninformative sentences by returning []. We refer to (pre-winnowing) SFT LLM CNs as "SFT LLM statements."

Following GH 1993, we approximate the distribution of unique CNs within the SFT LLM statements using a beta-binomial distribution. Let p_i be the probability that a block of b LLM

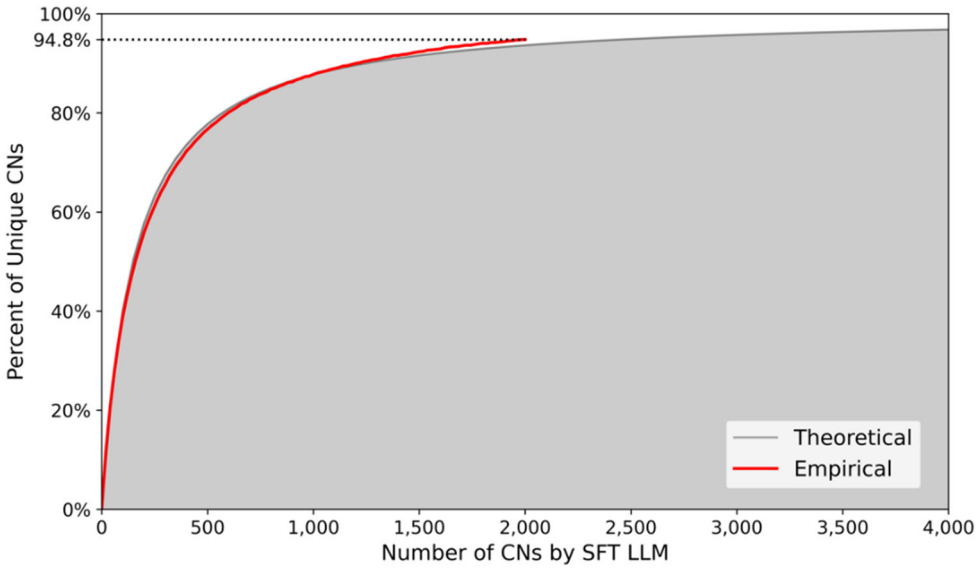
CNs contains a (final) CN i. For each of 2,000 SFT LLM statements, our data indicates which of the 117 final CNs, if any, were identified from it. We obtain an estimate of the p_i's by repeatedly sampling with replacement blocks of b SFT LLM statements and noting whether CN i was identified.

Using standard maximum likelihood estimation (MLE), we infer the parameters (α and β) of a beta distribution for the p_i's. The expected probability (E_n) of observing a CN from n blocks is given by Equation 1.

$$E_n = 1 - \frac{\Gamma(n+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\beta)} \quad (1)$$

In our data, $\alpha = 1.054$ and $\beta = 3.133$ for $b = 50$. Figure 6 plots E_n (grey area) including an extrapolation to 4,000 SFT LLM statements. Figure 6 also plots the average observed (resampled) probability p_i for up to 2,000 SFT LLM statements (red line). The indicated 94.8% is the resampling estimate of the number of CNs identified with 2,000 SFT LLM statements. The distribution of the p_i's has a beta-distribution shape and the beta-binomial model provides a reasonable fit to the observed data. The beta-binomial model suggests that 4,000 pre-winnowing SFT LLM statements would have contained, on average, approximately 96.8% of the wood-stain CNs. Interestingly, in this application, 1,000 SFT LLM statements would have contained approximately 87.7% of the wood stain CNs. We obtain similar results for $b = 1, 10, \text{and } 20$.

Figure 6: Customer Needs Extracted as a Function of the Number of Online Reviews



Characteristics of the SFT LLM Customer Needs

We asked our research partner to compare the characteristics of CNs extracted by professional analysts vs. CNs extracted by the SFT LLM. Using a consensus process, our research partner identified characteristics of CNs that are relevant to applications: (1) functional vs. emotional, (2) universal vs. niche, and (3) fleeting vs. enduring. Both functional and emotional CNs are valuable for product management and marketing – an effective VOC study identifies the right balance for the product or service being studied. Universal vs. niche also depends on the application. Niche is important when the market is highly segmented. Finally, both fleeting and enduring CNs are valuable for understanding customer experience. Fleeting needs occur at a specific moment in time (e.g., *can easily change a flat tire*), while enduring needs always exist (e.g., *able to check my tire pressure while driving*).

Professional analysts, blind to the source of identified CNs judged that two sets of CNs on the three characteristics. The SFT LLM was able to identify emotional CNs—a result that was

not fully expected. Indeed, the SFT LLM identified CNs that were more balanced—68% of the SFT-LLM-identified CNs were emotional vs. 83% for the analyst-identified CNs ($p = 0.07$). We do not have ground truth for the right percentage of emotional CNs, but, at minimum, we can say that the SFT LLM has the ability to identify emotional CNs. On average, there were no statistical differences in universal vs. niche ($p = 0.71$) or fleeting vs. enduring ($p = 0.91$).

Summary of SFT LLMs versus Human Analysts (Studies 1 & 2)

An SFT LLM identifies CNs as well or better than professional human analysts using the same corpus. It does not hallucinate any more than human analysts. A Base LLM is not sufficient. Given current technology, firms should not blindly use off-the-shelf LLMs. Instead, we find that just over 1,000 examples of CNs with verbatims are sufficient for the supervised fine-tuning. Finally, the SFT LLM scales well to the entire corpora and is not subject to fatigue. By using an SFT LLM to augment voice-of-the-customer applications, analysts can focus on higher value-added tasks and (likely) provide better strategic direction to product development, product management, and marketing strategy.

Additional Voice-of-the-Customer Applications

We report findings from two additional applications of the LLMs for identifying CNs. These applications provide valuable insights, recognizing that less complete evaluative data are available for each.

Oral Care

Following the procedures of Study 1, we applied both the Base LLM and the SFT LLM to extract CNs from the oral care reviews analyzed by TH 2019. The Base LLM and SFT LLM

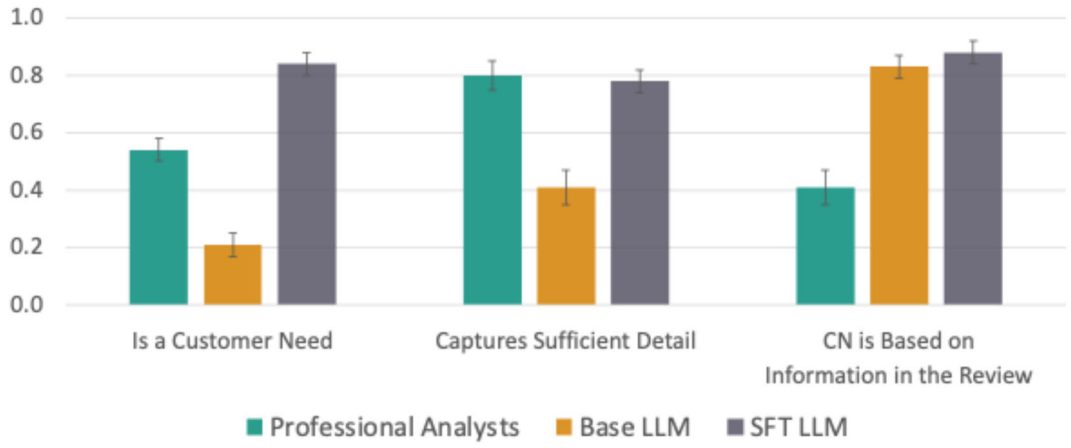
models are the same as in the wood stain application (without additional finetuning).

By necessity, the comparison among methods is slightly different between the wood stain and oral care. First, the oral care professional VOC application is substantially older than the wood stain application – the industry practice might have improved over time. Second, the samples were constructed slightly differently. The comparison for wood stain is based on 90 verbatims used in the original application, 30 additional informative reviews, and 30 uninformative reviews. For oral care, data construction started from the 86 CNs identified in the final affinity diagram for a for-client interview-based VOC study. The firm then reviewed UGC sentences and, for each sentence, identified all CNs relevant to this sentence. These differences affect the professional-analyst baseline, but not the Base LLM and SFT LLM extractions.

In Figure 7, we compare the performance of professional analysts, the Base LLM, and the SFT LLM in the oral care category. Despite the differences between the wood stain and oral care studies, the qualitative implications of the comparisons are the same. The SFT LLM is substantially and significantly better than the Base LLM on "is a customer need" ($p < 0.01$) and "sufficiently specific" ($p < 0.01$), and it is slightly better but not statistically significant on "follows from a verbatim" ($p = 0.09$). Furthermore, the SFT LLM is at least as effective as professional analysts and does substantially better on "is a customer need" ($p < 0.01$) and "follows from a verbatim ($p < 0.01$). The analysts do less well in oral care than wood stains, likely due to differences in the way in which the analyst baseline was created.

Importantly, similar to the wood stain analysis, the SFT LLM appears to be at least as accurate as professional analysts for extracting CNs. Finetuning is managerially and strategically relevant. The Base LLM does not appear to be sufficient.

Figure 7: Comparison of Professional Analysts and LLMs in the Oral Care Category



Applying the SFT LLM to Experiential Interviews

The applications in wood stain and oral care product categories used user-generated content (UGC) as source verbatims. Although the use of UGC is growing in VOC applications, most applications continue to rely on experiential interviews with customers. Our research partner assisted a professional product-development organization to interview academic thought leaders, researchers, society administrators, industry members, and potential industry members to understand customer needs for services provided by the organization. The professional analysts then reviewed the transcripts to identify CNs.

We obtained transcripts from all 20 customer interviews. After automatically parsing the transcripts into groups of semantically-connected sentences, we used an SFT LLM to extract CNs. The SFT LLM identified CNs that were judged to be legitimate CNs and to provide insight and value. The CNs identified by the two methods (SFT LLM and professional analysts) did not vary on functional vs. emotional or fleeting vs. enduring, but the SFT LLM was better at identifying niche CNs – an important characteristic given the heterogeneity in the interviewed customers.

There was strong, but not perfect, agreement among CNs identified by the SFT LLM and by professional analysts—69% of the primary CNs and 67% of the secondary CNs were identified by both the SFT LLM and professional analysts. The total shares of CNs identified by the SFT LLM were 87.5% and 88.9%, respectively for the primary and secondary CNs. Professional analysts identified 81.3% and 77.8%, respectively of the primary and secondary CNs. Although the SFT LLM did slightly better than professional analysts, the unique CNs identified by each method suggest that both methods added value in this interview-based application.

Summary, Limitations, and Future Research

In three applications, SFT LLMs effectively extract CNs from unstructured online reviews and experiential interviews. The SFT LLM CNs are coherent, follow professional standards, capture the underlying customer benefits at an abstract conceptual level, and are not hallucinations. The SFT LLM learned the task using training examples from ten VOC studies and could perform on par or better than professional analysts in new categories not in the training data. The SFT LLM was efficient and scaled well.

The result is surprising. Traditionally, professional communities in product development, product management, and marketing have emphasized the role of empathy in formulating CNs. By deeply understanding the customer experience and "jobs to be done," human analysts formulate the underlying CNs to spark customer insight and guide innovation. This task seems fundamentally human-centric and less susceptible to automation.

On the other hand, *a priori,* the need for finetuning is not obvious. LLMs are often human-like in many unstructured tasks. It would not have been surprising if the Base LLM performed well.

Why can SFT LLMs extract CNs so well? Foundational LLMs are trained on sentence completion tasks, which enables them to extract reasonable sentences and paraphrase well. During finetuning, we further calibrate the model by providing examples of how past VOC studies "paraphrase" customer reviews and interviews into CNs. The SFT LLM "learns" to imitate the paraphrasing exercise using professional standards. While the SFT LLM does not fundamentally understand the customer experience, the SFT LLM can talk about the customer experience. Without finetuning, an LLM does not seem to learn the specific task.

SFT LLM training resembles human learning in a professional environment. When a new analyst joins a firm, the analyst typically receives training explaining the differences between customer needs, solutions, opinions, and targets. Real-application trial-and-error experience imparts expertise. Experienced colleagues provide feedback and examples of professionally extracted CNs. This training can lead to new-product-development-professional certification (https://www.pdma.org/page/NPDP_certification). The training is designed to generalize. A product manager or new-product-development professional may have limited experience in the product category. Instead, they are trained to listen to the customers and elicit CNs by following well-honed procedures. The supervised finetuning approach mimics this training.

Limitations and Contributions

LLMs are advancing rapidly. They perform as well as human analysts on a variety of tasks, although not all tasks. We documented an important task in marketing that requires supervised finetuning. Improvements in LLMs should only widen the gap between human analysts and LLMs for the extraction of CNs. The Base LLM we tested was not sufficient to identify CNs. Extracting CNs is a very specific task; LLMs will likely continue to require finetuning to learn industry

standards.

LLMs are ubiquitous and supervised finetuning is well-established. Our contribution is assembling the training data and undertaking a careful evaluation of an SFT LLM with professional analysts. Our methods should have wide applicability. Voice-of-the-customer applications are widespread and have proven useful in many situations. Many firms are likely to have CN training data, which makes the methods readily deployable. On the other hand, evaluation by professional analysts is expensive, time-consuming, and requires commitment. Perhaps our careful and unbiased evaluation will encourage broad adoption.

Challenges Yet to be Met

SFT LLMs appear up to the job of extracting CNs. But there are many steps in the voice-of-the-customer that remain human-centric. We have not been able to automate either winnowing or affinity diagrams. CNs with similar meanings can be phrased very differently. Clustering embedding of source content, topic models, and keyword searches have been judged by professionals not to provide good structure. However, we are optimistic. CNs identified from SFT LLMs might be sufficiently standardized so that machine-based clustering methods become more effective for these tasks.

SFT LLMs extract CNS, but do not prioritize CNs. Researchers have attempted to use frequency, star-labeled ratings, and sentiment as indicators of importance, but such indicators are at best weakly correlated with CN importance and at worst counterproductive (GH 1993, TH 2019). Prioritization by machine learning remains elusive.

Finally, as we learn to automate more aspects of voice-of-the-customer applications, we can imagine a world where CNs and the fulfillment of CNs are monitored automatically and continuously to better manage the product, service, or brand. There are many steps between the

current state-of-the-art and that blue-sky goal, but automating CN extraction is a valuable step along that path.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida et al. (2023), "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774*.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura (2025), "AI-Human Hybrids for Marketing Research: Leveraging LLMs as Collaborators." forthcoming *Journal of Marketing*

Brand, James, Ayelet Israeli, and Donald Ngwe (2023), "Using LLMs for market research." *Available at SSRN* 23-062.

Brown, Tom B. (2020), "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165*.

Brown, Shona L., and Kathleen M. Eisenhardt. "Product development: Past research, present findings, and future directions." *Academy of management review* 20, no. 2 (1995): 343-378.

Burchill, Gary, and Christina Hepner Brodie (1997), *Voices into Choices: Acting on the Voice of the Customer* (Madison, WI: Joiner Associates, Inc)

Büschken, Joachim and Greg M. Allenby (2016), "Sentence-Based Text Analysis for Customer Reviews," *Marketing Science*, 35 (6), 953–75.

Cayla, Julien, Robin Beers, and Eric Arnould (2014), Stories that deliver business insights. *MIT Sloan Management Review* 55(2):54-62.

Chiang, Wei-Lin, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng 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.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. (2017), "Deep reinforcement learning from human preferences." *Advances in neural information processing systems* 30.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li et al. (2024), "Scaling instruction-finetuned language models." *Journal of Machine Learning Research* 25, no. 70, 1-53.

Ding, Min, John Hauser, Songting Dong, Daria Dzyabura, Zhilin Yang, Chenting Su, and Steven Gaskin (2011), "Unstructured Direct Elicitation of Decision Rules," *Journal of Marketing Research*, 48, (February), 116-127.

Dong, Guanting, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou (2023), "How abilities in large language models are affected by supervised fine-tuning data composition." *arXiv preprint*

arXiv:2310.05492.

Dong, Songting (2024), "Leveraging LLMs for Unstructured Direct Elicitation of Decision Rules." *Customer Needs and Solutions* 11:1.

Gao, Luyu, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig (2023), "Pal: Program-aided language models." In *International Conference on Machine Learning*, pp. 10764-10799. PMLR.

Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour (2024), "Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina." *arXiv preprint arXiv:2410.19599*.

Griffin, Abbie (2004), "Obtaining Customer Needs for Product Development" in Kenneth B. Kahn *The PDMA Handbook of New Product Development, 2E* (John Wiley & Sons, Inc. 211-227.

Griffin, Abbie and John R. Hauser (1993), The voice of the customer. *Marketing Science*. 12(1):1-27.

Gupta, Sunil (2020), "Are You Really Innovating Around Your Customers' Needs?," *Harvard Business Review*, October 1:1-5.

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

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

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

Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser et al. (2023), "ChatGPT for good? On opportunities and challenges of large language models for education." *Learning and individual differences* 103: 102274.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao, and Pablo Arredondo (2024), "Gpt-4 passes the bar exam." *Philosophical Transactions of the Royal Society A* 382, no. 2270: 20230254.

Lee, Thomas Y. and Eric T. Bradlow (2011), Automated marketing research using online customer reviews. *Journal of Marketing Research*. 48(5), 881-894.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. (2020), "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33: 9459-9474.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (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.

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

Min, Sewon, 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?." *arXiv preprint arXiv:2202.12837* .

Mitchell, John C. (2016) What is a customer need? *Pragmatic Marketing*.

Moor, Michael, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar (2023), "Foundation models for generalist medical artificial intelligence." *Nature* 616, no. 7956: 259-265.

Mullappilly, Sahal Shaji, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan (2023), "Arabic Mini-ClimateGPT: A Climate Change and Sustainability Tailored Arabic LLM." *arXiv preprint arXiv:2312.09366* .

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), “Mine Your Own Business: Market Structure Surveillance Through Text Mining,” *Marketing Science*, 31 (3): 521-543.

Qiu, Liying, Param Vir Singh, and Kannan Srinivasan (2023), "How Much Should We Trust LLM Results for Marketing Research?." *Available at SSRN 4526072*.

Rawte, Vipula, Amit Sheth, and Amitava Das (2023), "A survey of hallucination in large foundation models." *arXiv preprint arXiv:2309.05922*.

Shen, Junhong, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolo Fusi (2024), "Tag-LLM: Repurposing General-Purpose LLMs for Specialized Domains." *arXiv preprint arXiv:2402.05140*.

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom (2024), "Toolformer: Language models can teach themselves to use tools." *Advances in Neural Information Processing Systems* 36.

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

Selinger, Evan (2024), “How to stop believable bots from duping us all,” Boston Globe K1 November 24.

Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting (2023), "Large language models in medicine." *Nature medicine* 29, no. 8: 1930-1940.

Timoshenko, Artem and John R. Hauser (2019), “Identifying Customer Needs from User-Generated Content,” *Marketing Science*, 38:1, 1–20.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. (2023), "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample (2023), "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jukoh Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kalser, and Illia Polosukhin (2017), "Attention is all you need." *Advances in Neural Information Processing Systems*.

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

Zhang, Zheng, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao (2023), "Balancing specialized and general skills in LLMs: The impact of modern tuning and data strategy." *arXiv preprint arXiv:2310.04945*.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhouhan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseht E. Gonzales, and Ion Stolen (2024), "Judging LLM-as-a-judge with mt-bench and chatbot arena." *Advances in Neural Information Processing Systems* 36.

Zhu, Deyao, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny (2023), "Minigpt-4: Enhancing vision-language understanding with advanced large language models." *arXiv preprint arXiv:2304.10592*.

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. Vicuna Finetuning and Inference

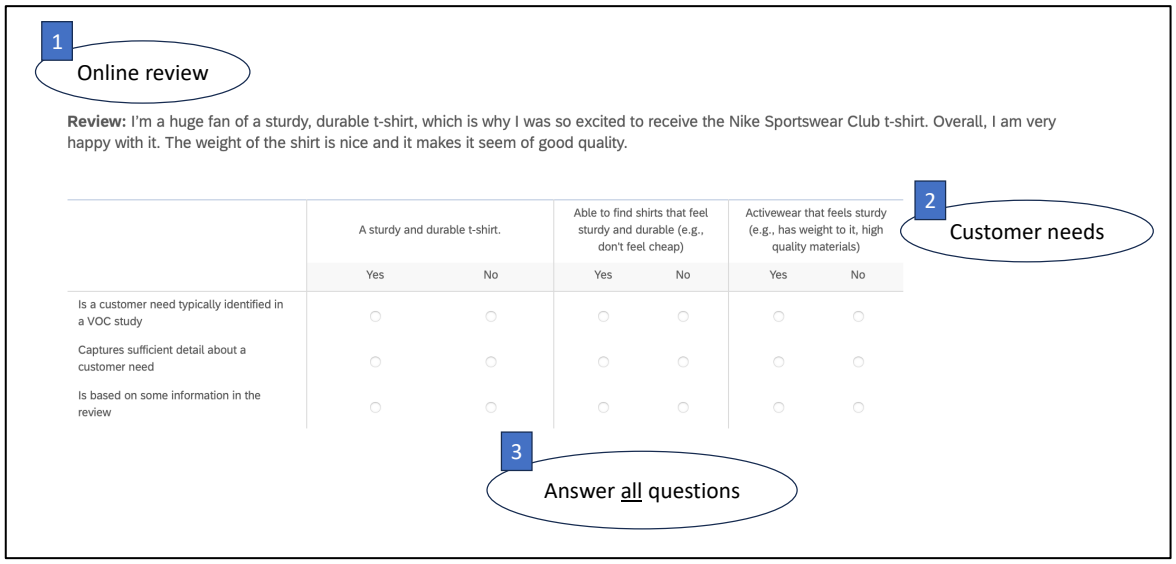
We finetuned Vicuna 13B using 4x Nvidia A100 PCIe on Lambda GPU Cloud. The system provides 40 GB VRAM, 120 vCPUs, 800 GB RAM, and 1 TB SSD storage. The finetuning process used bf16 precision without quantization, running for 6 epochs with a per-device batch size of 2 for training and 8 for evaluation. It employed a gradient accumulation of 4, a learning rate of 2e-5, cosine scheduling, and a maximum sequence length of 1024. Finetuning 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 C. 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 extracts 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 extraction 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 D. Detailed Instructions in Study 1



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 extract this customer need from the review.

Appendix E. Results for Study 1 Disaggregated by the Type of Source Material

	Verbatim			**Informative**			**Uninformative**		
	Human Analyst	Base LLM	SFT LLM	Human Analyst	Base LLM	SFT LLM	Human Analyst	Base LLM	SFT LLM
Is Customer Need	0.70 (0.23)	0.40 (0.26)	0.86 (0.21)	0.79 (0.27)	0.33 (0.27)	0.84 (0.22)	0.84 (0.24)	0.31 (0.24)	0.87 (0.24)
Sufficiently Specific	0.88 (0.19)	0.76 (0.33)	0.85 (0.25)	0.77 (0.25)	0.66 (0.32)	0.87 (0.18)	0.74 (0.28)	0.66 (0.34)	0.77 (0.25)
Follows from a Verbatim	0.84 (0.22)	0.86 (0.24)	0.92 (0.16)	0.06 (0.15)	0.88 (0.20)	0.90 (0.20)	0.04 (0.11)	0.52 (0.41)	0.76 (0.26)

We highlight two observations. First, on the dimension "Follows from a Verbatim," the performance of the human-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 other analyst-extracted CNs for these reviews from the original VOC study.

Second, we observe that the performance of the LLMs on the "Follow from a Verbatim" dimension is lower for *uninformative* reviews than for *informative* reviews and *verbatims*. Recall that analysts found these reviews uninformative in the professional voice-of-the-customer study. Our design substituted other CNs, so as expected, evaluators agreed that the analyst-extracted CNs are indeed CNs. The SFT LLM was able to extract CNs for many of the analyst-designated

uninformative reviews, suggesting that the SFT LLM is more efficient than analysts at identifying CNs. The Base LLM extracted some CNs, but not as many as the SFT LLM. The Base LLM shows evidence of hallucination in the “follows from a verbatim” question, particularly for uninformative reviews. The SFT LLM is more robust.