The BRAGE Benchmark: Evaluating Zero-shot Learning Capabilities of Large Language Models for Norwegian Customer Service Dialogues

Mike Riess
Research and Innovation
Telenor Group
Oslo, Norway
mike.riess@telenor.com

Tollef Emil Jørgensen
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
tollef.jorgensen@ntnu.no

Abstract

This study explores the capabilities of open-weight Large Language Models in a zero-shot learning setting, testing their ability to classify the content of customer service dialogues in Norwegian from a single instruction, named the BRAGE benchmark. By comparing results against widely used downstream tasks such as question-answering and named entity recognition, we find that (1) specific instruction models greatly exceed base models on the benchmark, (2) both English and multilingual instruction models outperform the tested Norwegian models of similar sizes, and (3) the difference between base and instruction models is less pronounced than in other generative tasks, suggesting that BRAGE is a challenging benchmark, requiring precise and generalizable instruction-tuning.

1 Introduction

Satisfied customers are critical to any telecommunications provider's long-term success and sustainability. An essential piece of this puzzle is to provide the best possible customer service once a problem has occurred and try to avoid any further negative experiences (PwC, 2018). Advances in Automatic Speech Recognition and text analysis methods have transformed customer service processes, enabling providers to gain aggregated insights from the large volume of daily calls. These insights allow the telecommunications provider to act quickly on issues that influence multiple customers in close to real-time. However, creating models capable of analyzing transcribed conversations remains challenging due to the technical expertise required and the time-intensive development process. Additionally, the distribution of the incoming calls may change over time due to concept drift (Riess, 2022), requiring frequent updating of models to maintain operational quality – thus increasing costs. In-context learning (Brown et al., 2020) and the ongoing efforts on adapting Large Language Models (LLMs) to lower-resource languages such as Norwegian (NORA AI, 2024) offer a promising solution to this problem.[1] This study explores the potential of open-weight LLMs to enable non-expert users to perform zero-shot content classification. To this end, we introduce *BRAGE*, a private benchmark designed to evaluate zero-shot classification of transcribed conversations in Norwegian. Using the same instructions provided to human annotators for creating ground truth labels, various LLMs are tasked with classifying the content of conversations between customers and customer service agents. To assess the feasibility of this approach, we evaluate base- and instruction-tuned open-weight LLMs, including pre-trained and fine-tuned Norwegian models. Given the sensitive nature of the data, BRAGE is a private benchmark. Aggregated results, however, are publicly shared to ensure transparency. Additionally, we[2] aim to facilitate academic collaboration by performing benchmark evaluations on BRAGE and sharing the results with the public when requested by researchers in the Nordics.
Code for the benchmark is available on GitHub.[3]

The Customer Service Process

The case process from which we create BRAGE is an analytics unit within a telecommunications provider, supporting customer service with insights on current calls. When customers contact the service center, calls are, if permitted, recorded

[1] Low-resource is relative to the amount of openly available resources for fine-tuning large language models, e.g., instruction-tuning. For Norwegian, this is largely limited to machine-translated datasets.

[2] Telenor Research and Innovation

[3] https://github.com/tnresearch/brage_2025

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 525–536
March 3-4, 2025 ©2025 University of Tartu Library

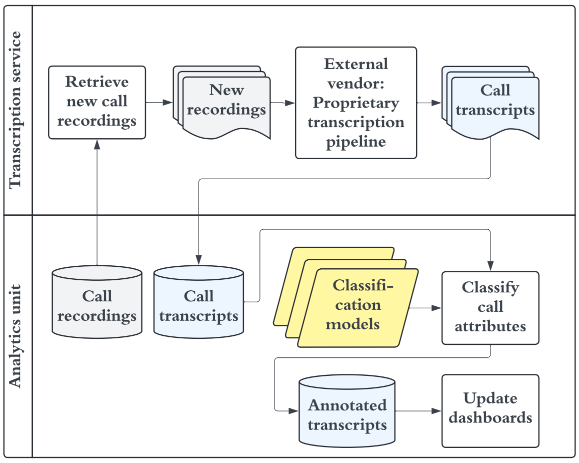

Figure 1: The call transcription and analysis process.

and subsequently transcribed.

Transcription is performed by a proprietary service from an external vendor, similar to *Whisper* (Radford et al., 2022). The transcripts are then processed and annotated using classification models, predicting business-relevant attributes. Figure 1 shows an overview of this process. In our benchmark, we modify the process by replacing the classification model(s) in Figure 1 with a single open-weight LLM, which is prompted using the *codebook* (annotation guidelines, see Forman and Damschroder, 2007) previously used by the human annotators.

Research Questions

We define the following research questions:

RQ1 How do open-weight Norwegian models compare regarding their performance on the BRAGE benchmark?

RQ2 How do these results compare and align with other downstream generative tasks for Norwegian?

To answer RQ1, we benchmark a set of open-weight LLMs, including Norwegian pre-trained and fine-tuned ones on BRAGE, and subsequently compare these results to other downstream evaluations from ScandEval (ScandEval, 2024) to answer RQ2.

2 Related work

In recent research and development of LLMs, it has become clear that these models can adapt to the context they are presented (Brown et al., 2020), to such a degree that they do not need further training to adapt to a particular task. This is also known as In-Context Learning (ICL) (Li et al., 2023), and can be done using a single instruction with no examples (*zero-shot*) or multiple examples at inference time (few/N-*shot*). ICL dramatically reduces the associated costs in evolving systems, as one no longer relies on expensive training pipelines to support shifts within data and business needs.

Open-weight models such as Llama (Dubey et al., 2024), Mistral (Jiang et al., 2023), and Gemma (Team et al., 2024) have proven to be competitive when considering the efficiency vs performance trade-off. “Smaller” models ($\leq$ 70B) have achieved impressive results across numerous tasks (Chiang et al., 2023; Wolfe et al., 2024), and we have reached a point where the performance gap between open-weight models and larger proprietary models is quickly diminishing. However, evaluating LLMs and quantifying this gap is incredibly difficult (Chandran et al., 2024; Biderman et al., 2024), e.g., because of benchmark data leaks (Xu et al., 2024). To circumvent this, LMSYS (Zheng et al., 2023b) developed an Elo-score system where users rank anonymous models – giving an idea of real-world performance. As of September 2024, several open-weight models ($\leq$ 70B) rank on par with much larger models.

Narrowing in on Scandinavian languages, ScandEval (Nielsen, 2023) is a tool for evaluating models on language-specific data for downstream tasks. Focusing on constrained generation with LLMs in this study, the Norwegian generative ScandEval benchmark is highly relevant, with 134 different models currently benchmarked across 9 different datasets.[4] Because our data is sensitive, the models must run locally to avoid transferring the data during inference. Open-weight models, particularly those explicitly trained on Norwegian data, are included for evaluation. While open models come with the advantage of enabling continued training, supervised fine-tuning (SFT), merging (Yang et al., 2024), and *jailbreaking* (Zou et al., 2023; Zhang et al., 2024) along with an ever-growing set of preference tuning techniques (Gao et al., 2024), the search space of optimization methods is so vast that finding an optimal approach is near impossible. Furthermore, Ghosh et al. (2024) found SFT to degrade

[4] As of October 2024. Visit `https://scandeval.com/norwegian-nlg/` for a full overview of Norwegian benchmark results.

knowledge and reduce output quality of the pre-trained models, with the same observations for using fine-tuned LLMs for telecommunications (Barnett et al., 2024). Moreover, research within out-of-domain generalizability in traditional machine learning and LLMs suggests that domain-specific training will reduce a model's performance (Wald et al., 2021; Yang et al., 2022; Yu et al., 2024). Mosbach et al. (2023) challenges this idea and performs thorough evaluations on the generalization of models for ICL and SFT in the parameter range of 125M to 30B. While Mosbach et al. find compelling results for the case of SFT, such that a smaller SFT can outperform higher-parameter models with ICL, these findings diminish as the model sizes grow, and ICL performs significantly better when evaluating in-domain than SFT for model sizes $\gtrapprox$ 7B.

3 Methodology

3.1 Data

The BRAGE benchmark consists of 300 transcribed Norwegian customer service phone calls from a telecommunication provider in its current version. Transcription is done by an external vendor with a proprietary algorithm. An internal validation of ten randomly sampled calls showed an average Word Error-Rate (WER) of 12.41% (overall) and 0.89% for business-critical terms like product names.

Annotation Each transcript has been annotated with several attributes related to each call, among those, the "product" attribute, which includes eight categories: *Annet* (Other), *Mobil* (Mobile), *Tjenester* (Services), *Bredbånd* (Broadband), *TV*, *Bredbånd-mobilt* (Broadband-Mobile), *E-post* (Email), *Forsikring* (Insurance). An internal analytics team developed the definitions for these categories in August 2023 and has since refined them multiple times to remove ambiguous categories. Experiments were conducted in October 2024. The 300 calls included in this study were randomly sampled and subsequently annotated by a Senior Analyst with 25+ years of domain experience (Customer Service) in two iterations: an initial annotation and a review two weeks after the initial annotation.

Class distribution The exact distribution of these categories cannot be shared due to business sensitivity. However, to provide the reader with an impression of the class imbalance, a randomly ordered overview is as follows: *Category 1 (*5.7%*), Category 2 (*13%*), Category 3 (*8%*), Category 4 (*7%*), Category 5 (*37%*), Category 6 (*9.3%*), Category 7 (*9%*), Category 8 (*11%*).* The expected accuracy of random guessing will thus be $\sum_{i=1}^{8} p_i^2 = 19.72\%$, where p_i is the individual class probability, while classifying all calls as *Category 5* will yield an accuracy of 37%. As this study evaluates *zero-shot* classification on a private test set, the models cannot use a zero-rate strategy (Devasena et al., 2011) or overfit to the majority class. To further account for class imbalance, we report our results with Macro-F1 and Matthews Correlation Coefficient in addition to Accuracy. 2 shows a modified example of a call transcript. This example shows the nature of the transcribed phone calls in the BRAGE benchmark, which is anonymized with _blacklist_, _number_-, and _name_ tokens. An English version can be found in Figure 7 in Appendix A.

3.2 Experiments

Each experiment consists of multiple runs, where a run represents a unique combination of a prompt and a model. For each run, we concatenate a zero-shot instruction (discussed in detail in section 3.3) with a truncated call transcript (the first 250 tokens), asking the model to determine which product the conversation is about. We utilize settings consistent with those employed in ScandEval (Nielsen, 2023), including a temperature of 0.0, a fixed seed value, and 10 iterations of bootstrap sampling for each run to ensure robustness.

The output space of the LLM is constrained to the valid product categories using Outlines (Willard and Louf, 2023) and Transformers (Wolf et al., 2020) for inference. We compare Pre-trained multilingual base models (P) Pre-trained base models in Norwegian Bokmål (PNB), Instruction-Tuned multilingual models (IT) and models Fine-tuned in Norwegian Bokmål (FNB). We use a single prompt for all models. Prompt formatting varies depending on the model type, but follow the guidelines in the model card from the respective authors. We use *ChatML* and *Alpaca* formats for instruction-tuned models, while base models receive prompts without any prior formatting. Upon completing all runs, we calculate aggregated metrics to evaluate the performance and outcomes of our experiments. To put our results

Hei, du snakker med _blacklist_. Jeg har gått over fra privat mobilabonnement til å få man dekket av jobben det skjedde for cirka to og en halv uke siden, og så ser jeg i nettbanken at de har en faktura som står til godkjenning for januar. Ja, ja. _number_ desember, da skulle liksom mobing av Mobilabonnementet det private avsluttes nå, så jeg lurte på nå kan jeg sjekke at faktureringen som har blitt riktig. _blacklist_. Ja, det skal hjelpe deg meg, kan jeg få ditt følge navn, fødselsdato og adresse. _blacklist_ _name_, _blacklist_ _blacklist_, _blacklist_ _blacklist_ _blacklist_ på _blacklist_ _blacklist_. Ja og postnummer A? _number_ _number_ _number_ _blacklist_. _blacklist_, men det var en telefon, ja, ja, du lurer på om faktum, altså du hadde en utestående faktura, sa du. Ja, i banken så legger jeg en faktura til godkjenning for januar. Ja. På _name_ sa du nei, den er betalt, fakturaen er betalt, ja. _number_ _number_ og _blacklist_ komma _blacklist_. Greit. Men når jeg lagt ned den fakturaen, så står det at det er for januar. Ja, da vil du få tilbake hele månedsprisen tilbake faktisk siden du abonnementet ditt ble endra, så ble endre da før januar starta, så får du alt jeg tilbake, ja, det er det, så du vil faktisk få tilbake skal vi sjå _number_ _number_ og og _blacklist_ det samme kontonummer du sist betalte med. Okay, ja, så da, da ble det på en måte avsluttet. Ja. Okay. Ja. På _number_ kroner, så da trekker vi fra den eller _blacklist_, men _blacklist_ ja. Det, ja. Den bare å avsette. Okay, ja for _blacklist_, så da, da får vi litt motta en sluttfaktura da. Greit, da glemmer jeg den fint ha det godt. Ha det, fint du.

Figure 2: Modified call example with similar quality as the transcripts in our dataset. The topic of this call is 'Mobile'. The terms _blacklist_, _name_ and _number_ are anonymized entities.

into perspective, we have retrieved the ScandEval scores of each model included in our study. The selected benchmarks cover the downstream tasks of named entity recognition (NorNE, Jørgensen et al., 2020), sentiment analysis (NoReC, Velldal et al., 2018), question-answering (NorQuAD, Ivanova et al., 2023) and commonsense reasoning (a truncated and machine-translated HellaSwag dataset Zellers et al., 2019, as implemented in ScandEval).

3.3 Prompting

The benchmark aims to assess LLMs' zero-shot performance on annotation tasks using human-equivalent instructions, evaluating the potential of automating the annotation task in a user-friendly manner. We, therefore, use the same guidelines when creating the ground truth. The only adaptation is to add a short introduction sentence and a final instruction for the model. The full prompt can be seen in Figure 3, which shows an anonymized version of this prompt. An English translation can be found in Figure 8 in Appendix A.

4 Results

4.1 RQ1: General Performance on the BRAGE benchmark

Looking at the right side of Figure 4, we observe that the variation across the base models is very low and that the average accuracy is around the same value as the expected accuracy of a random guess (19.72%). In stark contrast, the instruct models (left side) vary much more and have higher average accuracy, fostering the hypothesis that instruction fine-tuning is essential for our zero-shot classification task. Looking at the best performing models in Table 1, we find that the English/Multilingual *Gemma2* models outperform any model explicitly pre-trained (PNB) and/or fine-tuned on Norwegian data (FNB).

The average accuracy of the *Gemma2* models (43.53%, 62.1%, 60.53% for $2B$, $9B$ and $27B$ versions, respectively) also exceed random guess and the zero-rate classification. Amongst the models pre-trained or fine-tuned on Norwegian Bokmål we observe that the best-performing model is *NorskGPT Mistral 7B* with an average accuracy of 30.5%.

4.2 RQ2: BRAGE Performance Compared to ScandEval

To put our BRAGE results into perspective, we have organized a selection of ScandEval benchmarks into a radar chart to visualize the differences. The radar chart shows the relative accuracy across five benchmarks, where each polygon represents a model, and the area of the polygon the model performance. Figure 5 shows the results for English and multilingual Base and Instruct models, whereas Figure 6 shows the Norwegian Instruct models (right) and their corresponding models used for fine-tuning (left).

Looking at Figure 5, *Gemma2 9B IT* stands out with the highest average BRAGE accuracy

```
Her kommer det en liste med produktkategorier hos _brand_:\n - Mobil: _brand_ tilbyr
mobilabonnementer med bred dekning, ulike datapakker og tilbud på siste telefonmodeller.
Kategorien inneholder også datapakker og SIM-kort.\n \n - Forsikring: _brand_ tilbyr
forsikring for mobiltelefoner, som dekker tap, tyveri og skade, samt andre
forsikringsprodukter gjennom samarbejdspartnere. Produktene er _service_ og _service_.
Kategorien inneholder også henvendelser relatert til forsikringssaker, som behandles i en
egen avdeling. Kategorien skal ikke inneholde _service_ _service_, som skal kategoriseres
som Tjenester.\n \n - Annet: Kategorien når produkter ikke er spesifikt oppgitt i
samtalen. Gjelder særlig ved samtaler som er brutte eller når kunden har ringt feil. I
disse samtalene blir det ikke snakket om verken produkttype eller abonnement. \n \n -
E-post: _brand_ leverer sikre og pålitelige e-posttjenester med funksjoner for personlig
og profesjonell bruk, inkludert spamfiltrering og god brukervennlighet.\n \n -
Bredbånd-mobilt: _brand_ mobile bredbåndstjenester gir rask internettilgang på farten,
eller installert på fast adresse med utvendig antenne. Kategorien inneholder produktene
_service_, _service_, _service_ og _service_.\n \n - Tjenester: _brand_ tilbyr digitale
tjenester slik som sikkerhetsløsninger og skytjenester. Eksempler på tjenester er
_service_, _service_, _service_, _service_, _service_, _service_. I kategorien finnes
også _brand_ _service_, samt Trejepartstjenester som bl.a. omfatter innholdstjenester
som _service_.\n \n - Bredbånd: _service_ gir pålitelig internett med ulike
hastighetsalternativer, kombinert med kundevennlig service og teknisk support. I
kategorien finnes _service_ og _service_.\n \n - TV: _brand_ TV-tjenester inkluderer et
utvalg av kanalpakker, strømmetjenester og muligheter for opptak, alt tilpasset kundens
underholdningsbehov. Sentralt er produktet _service_, som er _brand_ TV-løsning.\n\n Her
er tekst fra en samtale mellom kundeservice og en kunde. Angi hvilken produktkategori
samtalen handler om, og svar kun med navnet på produktkategorien:\n <transcript>
```

Figure 3: Anonymized version of the prompt used. The text in **bold blue** is the prompt instruction added to the original guidelines used by the annotators, and <**transcript**> indicate where the conversation transcript is inserted.

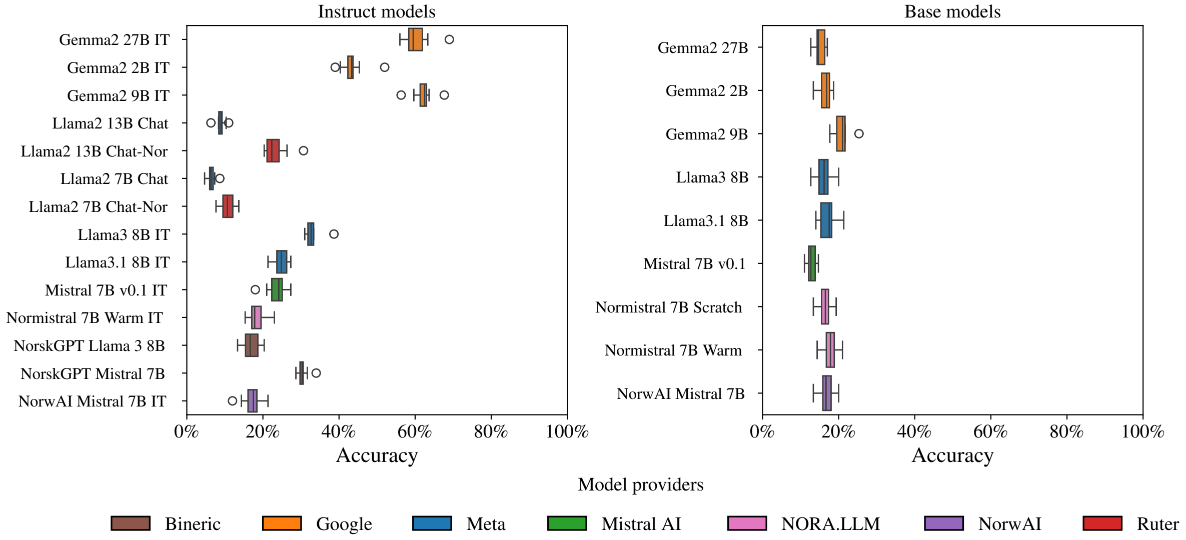

Figure 4: Comparison of accuracy distributions for instruction and base models. The color groupings separate them according to their respective provider/organization.

Benchmark /Metric(s)	**BRAGE**			**ScandEval** NorNE-nb	NoReC	NorQuAD	HellaSwag
Model	**Accuracy**	**Macro F1**	**MCC**	**Micro F1**	**Macro F1**	**F1**	**Accuracy**
Category: IT							
Gemma2 27B IT	60.53 ± 2.31	50.56 ± 1.54	54.27 ± 2.13	56.75 ± 3.04	**78.63 ± 0.96**	**73.41 ± 1.61**	**77.92 ± 1.72**
Gemma2 2B IT	43.53 ± 2.15	32.62 ± 1.50	33.25 ± 1.87	28.77 ± 2.22	63.18 ± 1.91	63.84 ± 1.50	49.42 ± 0.79
Gemma2 9B IT	**62.10 ± 1.80**	**53.50 ± 1.50**	**55.13 ± 1.91**	44.91 ± 3.62	73.45 ± 0.94	70.14 ± 1.53	75.79 ± 1.47
Llama2 13B Chat	8.93 ± 0.77	6.12 ± 0.85	-0.69 ± 1.36	40.40 ± 2.79	57.45 ± 3.77	69.24 ± 2.68	41.00 ± 1.40
Llama2 7B Chat	6.40 ± 0.71	2.69 ± 0.41	0.08 ± 1.32	38.59 ± 2.84	57.09 ± 3.80	61.99 ± 2.34	31.84 ± 1.05
Llama3 8B IT	33.03 ± 1.34	26.44 ± 1.23	25.38 ± 1.67	65.57 ± 2.39	65.69 ± 3.50	69.90 ± 3.17	45.85 ± 1.93
Llama3.1 8B IT	24.87 ± 1.13	21.80 ± 1.36	16.38 ± 1.93	**71.87 ± 0.97**	71.58 ± 0.90	70.96 ± 3.00	54.03 ± 0.82
Mistral 7B v0.1 IT	23.60 ± 1.74	17.33 ± 1.81	8.86 ± 2.33	34.52 ± 1.17	60.88 ± 1.36	63.67 ± 2.98	35.89 ± 1.06
Category: IT + FNB							
Llama2 13B Chat-Nor	23.27 ± 1.98	19.79 ± 1.10	14.59 ± 2.05	47.74 ± 2.83	58.47 ± 3.79	65.76 ± 3.07	41.29 ± 1.19
Llama2 7B Chat-Nor	10.80 ± 1.14	8.69 ± 1.29	2.80 ± 1.51	20.44 ± 2.47	23.50 ± 3.03	50.11 ± 1.80	24.48 ± 0.70
NorskGPT Llama 3 8B	16.87 ± 1.51	15.14 ± 1.50	6.14 ± 1.90	**60.25 ± 3.14**	61.42 ± 3.56	**74.57 ± 2.20**	59.11 ± 2.44
NorskGPT Mistral 7B	**30.50 ± 0.91**	**26.89 ± 1.05**	**22.54 ± 1.47**	47.72 ± 3.74	**70.81 ± 1.30**	74.38 ± 3.92	**60.59 ± 1.18**
Category: P							
Gemma2 27B	15.07 ± 0.85	13.97 ± 0.96	6.08 ± 1.06	43.06 ± 1.89	**76.14 ± 1.68**	**80.21 ± 4.49**	**63.55 ± 4.76**
Gemma2 2B	16.43 ± 1.00	13.08 ± 1.02	4.98 ± 1.34	21.28 ± 2.58	47.91 ± 2.11	63.31 ± 3.73	28.89 ± 1.54
Gemma2 9B	**20.80 ± 1.31**	**17.02 ± 1.20**	**7.51 ± 1.41**	34.62 ± 1.80	75.53 ± 0.73	72.99 ± 3.16	63.52 ± 3.49
Llama3 8B	16.23 ± 1.35	14.16 ± 1.18	3.38 ± 1.57	47.65 ± 2.94	66.15 ± 1.44	74.98 ± 3.70	42.47 ± 2.74
Llama3.1 8B	17.07 ± 1.43	14.48 ± 1.49	4.36 ± 1.70	**53.50 ± 3.27**	68.71 ± 1.01	75.98 ± 2.62	46.84 ± 1.59
Mistral 7B v0.1	12.87 ± 0.74	10.31 ± 0.73	-0.20 ± 0.62	43.55 ± 2.21	64.53 ± 3.71	70.86 ± 2.79	32.43 ± 2.67
Category: PNB							
Normistral 7B Scratch	16.47 ± 1.09	**13.89 ± 1.41**	1.99 ± 1.51	15.44 ± 5.52	36.85 ± 2.01	38.93 ± 2.59	24.84 ± 0.71
Normistral 7B Warm	**17.83 ± 1.29**	13.76 ± 1.27	**2.31 ± 1.54**	**31.45 ± 1.88**	45.30 ± 3.46	61.85 ± 3.07	25.00 ± 0.83
NorwAI Mistral 7B	16.83 ± 1.19	12.55 ± 1.24	1.22 ± 1.46	20.45 ± 2.65	**65.98 ± 2.95**	**68.04 ± 5.37**	**27.82 ± 1.56**

Table 1: Aggregated performance metrics of BRAGE and a selection of ScandEval results from 10 runs. BRAGE performance is reported in Accuracy, Macro F1, and Mathews Correlation Coefficient (MCC), each with their corresponding ± 95% confidence intervals (CI). ScandEval results include individual scores per benchmark and confidence intervals (see Nielsen, 2023; Nielsen et al., 2024). Model category abbreviations: Pre-trained on Norwegian Bokmål (PNB), Fine-tuned on Norwegian Bokmål (FNB), Pre-trained (P), Instruction-tuned (IT). The highest scores for each category are boldfaced.

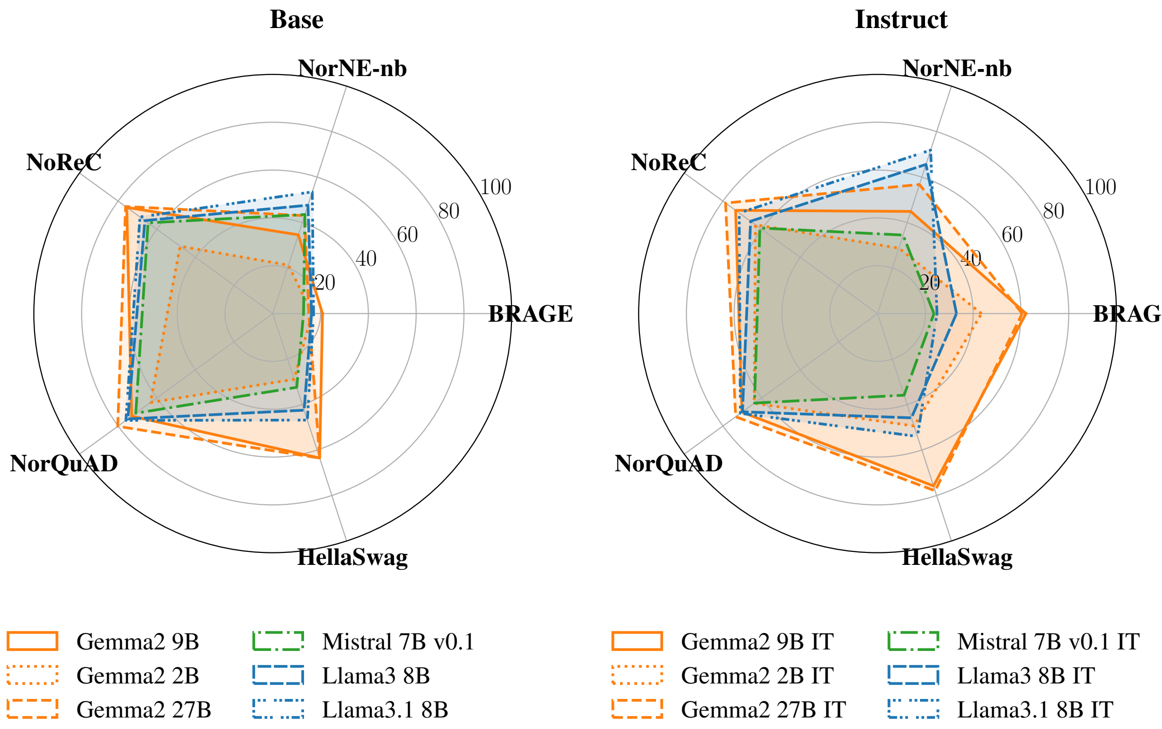

Figure 5: English/Multilingual models. Left: base models. Right: instruction-tuned. Radar chart of scores on selected ScandEval datasets and BRAGE results.

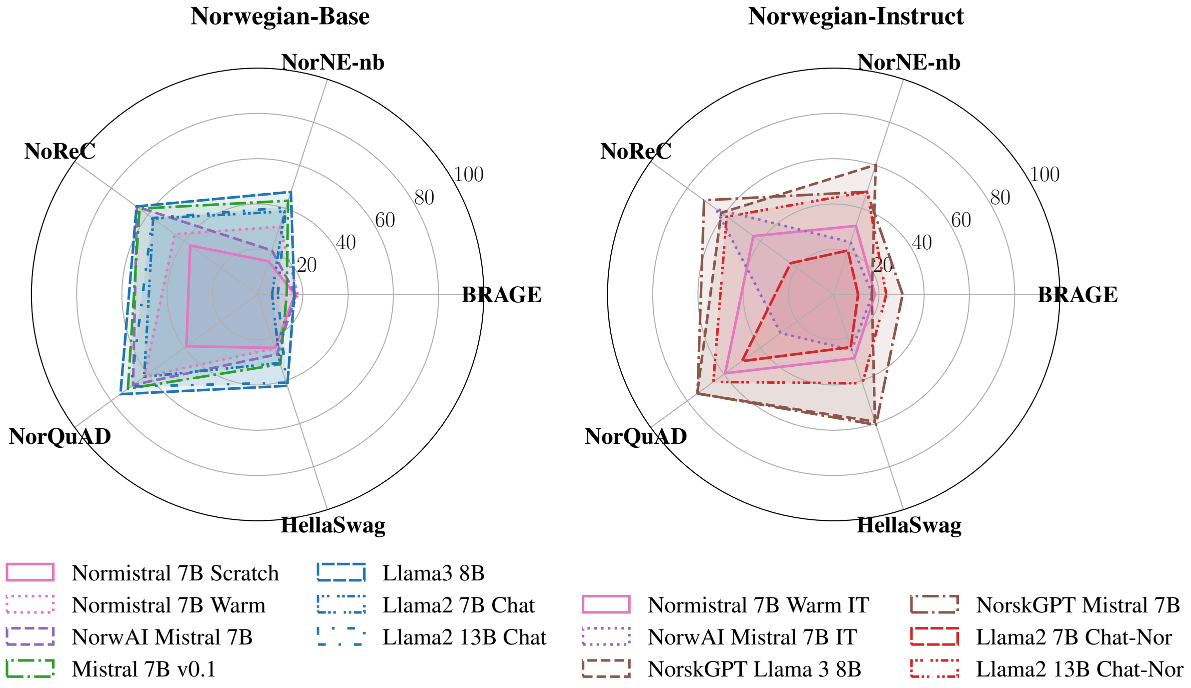

Figure 6: Norwegian models (and their corresponding base models used for fine-tuning). Left: base models. Right: fine-tuned on Norwegian. Radar chart of scores on selected ScandEval datasets and BRAGE results.

of 62.1%. Although somewhat below the 27B model for the ScandEval benchmarks, these findings are mostly consistent, except for the named entity task (NorNE-nb), where the Llama-models (e.g. *Llama3.1 8B IT*) surpass all other models, with a micro-F1 of 71.87 compared to 56.75 of the *Gemma2 27B IT*.

Moving on to the Norwegian models in Figure 6, pre-training (PNB) and fine-tuning (FNB) tend to lag behind the models fine-tuned on non-public data sources (IT + FNB). The architectural choices, and especially the fine-tuning procedures, seem to have a much higher importance for the BRAGE benchmark, as well as for HellaSwag (commonsense reasoning tasks), where we see a close relationship in terms of performance deltas: high scores on HellaSwag indicates high scores for BRAGE. In contrast, high scores on NorNE, for example, do not follow this pattern.

5 Discussion

Bigger is not Always Better

While larger models tend to get better results overall, we observe that *Gemma2 9B IT*, through its knowledge distillation training process (Team et al., 2024), approximates (and even outperforms) the larger version at 27B parameters, which is in alignment with other public benchmarks, such as *open llm leaderboard*.[5] Moreover, the 8B Llama models perform well on several tasks, especially for named entity recognition, outscoring the larger models. These effects will likely become more prominent as smaller models are trained through knowledge distillation and fine-tuned on domain-specific tasks.

Instruction Tuning

Good results for instruction-tuned (IT) models on other benchmarks did not necessarily translate to BRAGE. We have noted the relation between HellaSwag for IT models, but the base models still achieve relatively high scores on all downstream tasks. In contrast, BRAGE requires specific fine-tuning to achieve good results, as exemplified by only the Gemma2 models reaching acceptable accuracy scores. This, too, is the case for the base model Mistral 7B v0.1 (score of 13.60) compared to 29.58 for the NorskGPT Mistral 7B, while increasing its HellaSwag score from 32.43 to 60.59, while other tasks remain nearly unchanged in comparison. The fine-tunes by NorwAI and NORA.LLM (PNB + FNB) have approximately equal scores and lower stability than the base models. Additionally, we believe some Norwegian models are fine-tuned using the presented datasets, which, in turn, results in poor generalizability. Note the high deviance by, e.g., *Norwai Mistral 7B (PNB)* scoring surprisingly high on NoReC and NorQuAD, but not NorNE. The opposite is the case for *Normistral 7B Warm (PNB)*.

Suggestions for Future Work

Evidently, modeling decisions, data, post-training fine-tuning, and alignment require extra attention. Few organizations share end-to-end details – besides the OLMo initiative (Groeneveld et al., 2024), and we are typically left with a higher-level view of potential improvements for future developments of LLMs. Based on our findings, the Gemma2 architecture seems suitable for most of our tests and public benchmarks, and we leave the following suggestions for language-specific LLM development in the post-training stage:

- Distillation to student distributions, keeping compute-optimal token counts in mind (Gu et al., 2023; Agarwal et al., 2024).
- Different reward setups through RLHF (Christiano et al., 2017) and other alignment procedures (Gao et al., 2024).
- Incorporating prompts from, e.g., LMSYS-CHAT-1M (Zheng et al., 2023a), with responses from larger teacher models.
- Studies on instruction formatting.

Business Perspectives

The potential value contribution of zero-shot LLM-based content classification to customer service operations is significant in terms of both user-friendliness and development time. However, our results suggest that the current performance level is not yet sufficient for full production deployment, suggesting a need for further research and development in this area.

Furthermore, the suggested approach relies on ground truth to assess model quality pre-production, only partially automating the content classification process. Finally, using highly resource-consuming LLMs for a task that can be

[5] `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`

solved using smaller, more energy-efficient models raises questions regarding sustainability and cost versus benefit (Rigutini et al., 2024).

6 Conclusion

We have presented BRAGE, a private zero-shot benchmark for classifying transcribed calls between customers and customer service. Based on these preliminary results, we observe that the task can be accomplished to a somewhat acceptable level using open-weight LLMs. Based on our results, we can conclude that this is a challenging benchmark and that instruction fine-tuned models generally perform better on this type of zero-shot task. Specifically, instruction fine-tuning (FNB) on a multilingual base model, in the case of *NorskGPT Mistral 7B*, was superior to any of the other Norwegian models on BRAGE. We, therefore, stress the importance of creating more open instruction datasets in Norwegian, as this might foster progress in zero-shot settings such as the BRAGE case. Surprisingly, we found that the *English-only* and significantly smaller *Gemma-2 2B IT* did better than any of the Norwegian models. These results may also apply to other European languages, especially those with a higher presence in multilingual training corpora, e.g., German and Spanish. We plan to expand this benchmark by adding new tasks as well as to include all of the Scandinavian languages.

7 Limitations

As these experiments were conducted on a real business case, relevant information, such as distribution details about our data, had to be left out due to its sensitivity. However, we hope BRAGE, as a private benchmark can still be a contribution to the academic community, when committing ourselves to share aggregated results with the public (keeping data private on local infrastructure) going forward. Our conclusions also remain limited by the amount of information publicly available on the models included in the study, we therefore specifically hope to see more published data concerning pre-training and instruction-tuning for the current and future research-funded models (e.g., by NORA.LLM and NorwAI).

8 Sustainability

We have tracked power consumption and estimated emissions for all experiments using CodeCarbon (Schmidt et al., 2021). Hardware: 4x RTX 3090 GPUs over 29.2 hrs, resulting in a total emission of $0.8394 \text{ kgCO}_2\text{e}$ given the energy mix in Oslo, Norway.

Acknowledgments

The authors would like to thank the team at the case company for their cooperation and contributions: Tom Arne Gjesdal Karlsen, Ole Jonas Liahagen, Stig Hodnebrog and Noman Bashir.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Scott Barnett, Zac Brannelly, Stefanus Kurniawan, and Sheng Wong. 2024. Fine-tuning or fine-failing? debunking performance myths in large language models. *arXiv preprint arXiv:2406.11201*.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. 2024. Private benchmarking to prevent contamination and improve comparative evaluation of llms. *arXiv preprint arXiv:2403.00393*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, 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.

C Lakshmi Devasena, T Sumathi, VV Gomathi, and M Hemalatha. 2011. Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface*, 1:5.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jane Forman and Laura Damschroder. 2007. Qualitative content analysis. In *Empirical methods for bioethics: A primer*, pages 39–62. Emerald Group Publishing Limited.

Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, Zhe Yang, Liang Chen, Helan Hu, Runxin Xu, Qingxiu Dong, Ce Zheng, et al. 2024. Towards a unified view of preference learning for large language models: A survey. *arXiv preprint arXiv:2409.02795*.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. 2024. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. Norquad: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168.

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

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. Norne: Annotating named entities for norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4547–4556.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks. *arXiv preprint arXiv:2406.13469*.

NORA AI. 2024. Big steps towards a norwegian answer to chatgpt. Accessed: 2024-10-03.

PwC. 2018. Experience is everything: Here's how to get it right. Consumer intelligence series, PricewaterhouseCoopers.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Mike Riess. 2022. Automating model management: a survey on metaheuristics for concept-drift adaptation. *Journal of Data, Information and Management*, 4:211–229.

Leonardo Rigutini, Achille Globo, Marco Stefanelli, Andrea Zugarini, Sinan Gultekin, and Marco Ernandes. 2024. Performance, energy consumption and costs: A comparative analysis of automatic text classification approaches in the legal domain. *International Journal on Natural Language Computing*.

ScandEval. 2024. Danish NLU. A Natural Language Understanding Benchmark.

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. 2021. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and Bill Howe. 2024. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, volume 35 of *FAccT '24*, page 1199–1210. ACM.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.

Han Yu, Jiashuo Liu, Xingxuan Zhang, Jiayun Wu, and Peng Cui. 2024. A survey on evaluation of out-of-distribution generalization. *arXiv preprint arXiv:2403.01874*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tianrong Zhang, Bochuan Cao, Yuanpu Cao, Lu Lin, Prasenjit Mitra, and Jinghui Chen. 2024. Wordgame: Efficient effective llm jailbreak via simultaneous obfuscation in query and response.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

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

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

A Translated Transcripts and Prompts

Figure 7 shows the translated call transcript, and Figure 8 shows the translated prompt template for transcripts.

Hi, you're speaking with _blacklist_. I have switched from private mobile subscription to having it covered by work, this happened about two and a half weeks ago, and then I see in the online bank that they have an invoice pending approval for January. Yes, yes. _number_ December, that's when mobing of mobile subscription, the private one, was supposed to be terminated, so I was wondering now can I check that the billing has been correct. _blacklist_. Yes, I'll help you with that, can I have your last name, date of birth and address. _blacklist_ _name_, _blacklist_ _blacklist_, _blacklist_ _blacklist_ _blacklist_ at _blacklist_ _blacklist_. Yes and postal code A? _number_ _number_ _number_ _blacklist_. _blacklist_, but it was a phone, yes, yes, you're wondering about the invoib, so you had an outstanding invoice, you said. Yes, in the bank I put an invoice pending approval for January. Yes. For _name_ you said no, it's paid, the invoice is paid, yes. _number_ _number_ and _blacklist_ comma _blacklist_. Okay. But when I put down that invoice, it says it's for January. Yes, then you will get back the entire monthly fee actually since your subscription was changed, it was changed before January started, so you get everything back, yes, that's it, so you will actually get back let's see _number_ _number_ and and _blacklist_ to the same account number you last paid with. Okay, yes, so then, then it was kind of terminated. Yes. Okay. Yes. For _number_ kroner, so then we deduct that or _blacklist_, but _blacklist_ yes. That, yes. Just set it aside. Okay, yes for _blacklist_, so then, then we'll receive a final invoice then. Alright, then I'll forget about that, fine goodbye. Goodbye, you too.

Figure 7: English translation of a modified call example with similar quality as the transcripts in our dataset. The topic of this call is 'Mobile'. The terms _blacklist_, _name_ and _number_ are anonymized entities.

Here comes a list of product categories at _brand_:\n - Mobile: _brand_ offers mobile subscriptions with broad coverage, various data packages and offers on latest phone models. The category also includes data packages and SIM cards.\n \n - Insurance: _brand_ offers insurance for mobile phones, covering loss, theft and damage, as well as other insurance products through collaboration partners. The products are _service_ and _service_. The category also includes inquiries related to insurance cases, which are handled in a separate department. The category should not include _service_ _service_, which should be categorized as Services.\n \n - Other: The category when products are not specifically mentioned in the conversation. Applies particularly to conversations that are broken or when the customer has dialed wrong. In these conversations, neither product type nor subscription is discussed.\n \n - Email: _brand_ delivers secure and reliable email services with features for personal and professional use, including spam filtering and good user-friendliness.\n \n - Broadband-mobile: _brand_ mobile broadband services provide fast internet access on the go, or installed at a fixed address with external antenna. The category contains the products _service_, _service_, _service_ and _service_.\n \n - Services: _brand_ offers digital services such as security solutions and cloud services. Examples of services are _service_, _service_, _service_, _service_, _service_, _service_. The category also includes _brand_ _service_, as well as Third-party services which include content services like _service_.\n \n - Broadband: _service_ provides reliable internet with various speed options, combined with customer-friendly service and technical support. The category includes _service_ and _service_.\n \n - TV: _brand_ TV services include a selection of channel packages, streaming services and recording options, all adapted to the customer's entertainment needs. Central is the product _service_, which is _brand_'s TV solution.**\n\n Here is text from a conversation between customer service and a customer. Indicate which product category the conversation is about, and respond only with the name of the product category:\n** <transcript>

Figure 8: English translation of the anonymized version of the prompt used. The text in **bold blue** is the prompt instruction added to the original guidelines used by the annotators, and <**transcript**> indicate where the conversation transcript is inserted.