Reformulate and Create, Don't Translate: Creating Natural Prompts for Underserved Languages

Annika Simonsen[1], **Mathias Stenlund**[1], **Lars Bungum**[2], **Marc Daníel Skipstað Volhardt**[1] **and Hafsteinn Einarsson**[1]
[1]University of Iceland, Sæmundargata 2, 102 Reykjavík, Iceland
[2]Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, 7034 Trondheim, Norway
{annika, hem29, mds, hafsteinne}@hi.is, lars.bungum@ntnu.no

Abstract

We present a methodology for creating high-quality instruction prompts for low-resource Germanic languages that addresses a critical challenge: small annotator pools risk producing datasets reflecting narrow individual interests rather than diverse user needs. In this work, native speakers reformulate existing English prompts from OpenAssistant or create entirely original prompts, adapting them to reflect local contexts and natural language patterns while preserving broad task and topic diversity. This approach produced high-quality prompt datasets totaling 6,950 prompts across seven Germanic languages (German, Dutch, Swedish, Norwegian Bokmål/Nynorsk, Danish, Icelandic and Faroese) with validated coverage of diverse tasks and topics. Blind evaluation demonstrates that human-reformulated prompts significantly outperform synthetically generated prompts in naturalness and comprehensibility, particularly for low-resource languages like Icelandic and Faroese. For the bigger Scandinavian language, Danish, the difference was less pronounced. The prompt dataset is released under an open-source license at https://huggingface.co/datasets/AnnikaSimonsen/TrustLLM-reformulation-prompts.

1. Introduction

Aligned multilingual Large Language Models (LLMs) require high-quality instruction-following data capturing diverse user intents and cultural contexts. However, instruction fine-tuning and RLHF (Ouyang et al., 2022) datasets are primarily in English or in other high-resource languages, creating capability gaps for low-resource language speakers and risking cultural misrepresentations in AI systems (Arora et al., 2023; Cao et al., 2023).

Existing approaches face fundamental tradeoffs. Human annotation ensures quality but is financially infeasible for LLM developers, while translation fails to capture cultural nuances (Hershcovich et al., 2022; Cao et al., 2024; Debess et al., 2025). Purely synthetic generation lacks diversity and inherits model biases (Chen et al., 2024; Yu et al., 2023). Germanic languages beyond English and German have virtually no human-written instruction fine-tuning data available. In Norwegian, instruction tuning has primarily relied on translated versions of the Alpaca dataset (Taori et al., 2023), such as those used in NorGLM, although a few human-written datasets have only recently been released (see Section 2).

Another challenge for small-scale manual annotation is the following: with few contributors, resulting prompts risk reflecting narrow individual interests rather than capturing diverse user needs. This annotator bias becomes severe for low-resource languages where large annotator pools are not possible. Question answering dataset development addressed similar challenges, such as Natural Questions (Kwiatkowski et al., 2019), TyDi QA (Clark et al., 2020), and Natural Questions in Icelandic (Snæbjarnarson and Einarsson, 2022) demonstrated that having native speakers reformulate diverse seed data produces datasets better capturing real-world language use. Our research question therefore is; can we apply similar principles to instruction data, transforming existing diverse English prompts into high-quality, culturally relevant prompts for low-resource languages while maintaining task and topic diversity?

We address these challenges through a native speaker-driven methodology that combines reformulation of diverse English prompts with original prompt creation (see Figure 1). Native speakers adapt OpenAssistant's (Köpf et al., 2023) English prompts for their linguistic communities or create entirely new prompts, preserving task diversity while ensuring cultural authenticity and natural language use.

Our work makes three primary contributions: (1) 5,950 human-written prompts across seven Germanic languages (Swedish, Norwegian Bokmål, Danish, Icelandic, Faroese, Dutch, and German) and 1,000 machine translated, human-validated prompts for Norwegian Nynorsk with validated task and topic taxonomy; (2) an open-source annotation platform for multilingual prompt reformulation and creation, adaptable to other language families; (3) empirical evidence through blind evaluation that human reformulation significantly outperforms synthetic generation for low-resource languages, particularly for Icelandic and Faroese.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10735–10749
11–16 May 2026.

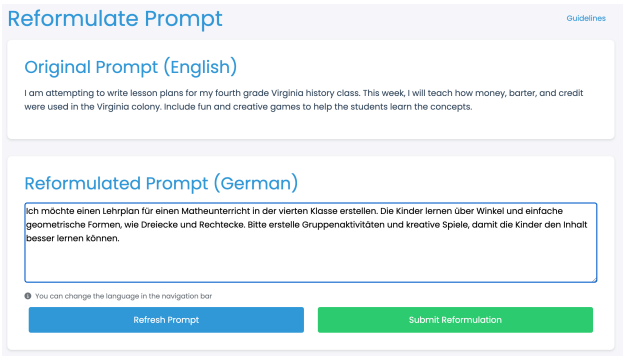

Figure 1: An example of the reformulation task in German. The English sentence is from the OASST2 dataset (Köpf et al., 2023). The German reads "I would like to create a lesson plan for a math class in fourth grade. The children learn about angles and simple geometric shapes, such as triangles and rectangles. Please create group activities and creative games so that the children can learn the content better.".

2. Related Work

In this Section, we review existing instruction datasets for Germanic languages and previous work on multilingual prompt creation methodologies, highlighting gaps that motivate our reformulation approach.

2.1. Instruction Datasets for Germanic Languages

English instruction fine-tuning datasets range from small curated collections like LIMA (1K samples (Zhou et al., 2023)) to massive compilations like H4 (10.8M rows (Lambert et al., 2023)), using human (Conover et al., 2023; Muennighoff et al., 2024), synthetic (Wang et al., 2024, 2022), or hybrid methods (Fang et al., 2024; Kim et al., 2022).

For non-English Germanic languages, resources are substantially sparser. Table 1 summarizes existing datasets, revealing substantial imbalances: while German, Dutch and Norwegian have some resources, languages like Swedish, Danish, Icelandic, and Faroese remain severely underserved.

Dataset	Lang.	N
German		
GermanQuAD (Möller et al., 2021)	DE	13k
OpenAssistant (Köpf et al., 2023)	DE	6.0k
MLQA (Lewis et al., 2020)	DE	5k
UltraChat^{T} (Patz, 2024)	DE	1k
Chatbot Arena (Zheng et al., 2023)	DE	676
Aya (Singh et al., 2024)	DE	241
Dutch		
Dutch SQuAD^{T} (Rouws et al., 2022)	NL	104k
Aya (Singh et al., 2024)	NL	1.7k
Chatbot Arena (Zheng et al., 2023)	NL	87
OpenAssistant (Köpf et al., 2023)	NL	72
Swedish		
Unnat. Instr.^{T} (Holmström and Doost., 2023)	SV	66k
gsm8k^{T} (Holmström and Doost., 2023)	SV	8k
Aya (Singh et al., 2024)	SV	1.3k
Chatbot Arena (Zheng et al., 2023)	SV	24
OpenAssistant (Köpf et al., 2023)	SV	1
Norwegian		
Mimir Instruction (de la Rosa et al., 2025)	NO	5k
NRK-Quiz-QA (Mikhailov et al., 2025)	NO	4.9k
NorQuAD (Ivanova et al., 2023)	NO	4.7k
NorOpenBookQA (Mikhailov et al., 2025)	NO	3.5k
NorCommonSenseQA (Mikhailov et al., 2025)	NO	1k
NorTruthfulQA (Mikhailov et al., 2025)	NO	1k
Chatbot Arena (Zheng et al., 2023)	NO	19
Danish		
SkoleGPT^{T} (Junge et al., 2024)	DA	21.6k
Aya (Singh et al., 2024)	DA	97
OpenAssistant (Köpf et al., 2023)	DA	44
Chatbot Arena (Zheng et al., 2023)	DA	10
Icelandic		
Chatbot Arena (Zheng et al., 2023)	IS	1
Multi-language datasets		
Scandi-QA† (Nielsen, 2023)	SV/NO/DA	7.8k
Belebele (Bandarkar et al., 2024)	ALL	900

Table 1: Existing instruction and QA datasets for Germanic languages, ordered by size within each language group. Unmarked datasets are human-written; †Hybrid (human questions, machine-translated context); ^{T}Machine-translated.

2.2. OpenAssistant

The OpenAssistant project (Köpf et al., 2023) was a global crowdsourcing campaign involving over 13,500 volunteers, creating the OpenAssistant Conversations (OASST1) dataset with 161,443 messages in 35 languages and 461,292 human quality ratings. The newer OASST2 version reveals severe imbalances: while English has 64,513 samples in the ready_for_export subset, German has 6,145, Dutch only 72, Danish 44, and Swedish merely 1 sample. This motivated our work on efficient methods for creating culturally appropriate instruction data.

While OpenAssistant's guidelines emphasize "diverse and challenging" inputs (Köpf et al., 2023), to our knowledge, no systematic analysis of task or topic distribution has been published. Our taxonomy analysis (Section 3.3) of our reformulated dataset reveals imbalances inherited from OASST2.

2.3. Instruction Taxonomy Frameworks

Effective categorization of instruction data requires taxonomies aligning with model capabilities and real-world usage. MT-Bench (Zheng et al., 2023) established an influential 8-category framework covering writing, roleplay, extraction, reasoning, math, and coding. The Databricks Dolly dataset (Conover et al., 2023) demonstrated practical operationalization by transforming InstructGPT's be-

havioral categories into crowdsourced annotations. Large-scale analysis of ChatGPT usage (Chatterji et al., 2025) revealed that 80% of real-world conversations cluster into three categories: Practical Guidance, Seeking Information, and Writing.

3. Methodology

We describe our approach in four stages: first, the prompt reformulation and creation process that enables native speakers to produce culturally authentic prompts; second, our platform implementation and contributor management strategy; third, our task and topic taxonomy design for systematic diversity assessment; and fourth, our evaluation methodology comparing human and synthetic prompt quality.

3.1. Prompt Collection: Reformulation and Creation

We developed a specialized annotation platform for reformulating OASST2 English prompts into Germanic languages. Using this annotation platform, our methodology encompasses two annotation tasks:

Prompt Reformulation Annotators receive English prompts from OASST2 and reformulate them in their native language. This process explicitly goes beyond translation; annotators may maintain the original intent or alter it while staying within the same task category. For example, a Swedish contributor might take an English prompt that asks for a recipe for red velvet cake and reformulate it to ask for a recipe for *prinsesstårta*. The goal is creating prompts that sound natural in the target language, incorporate culturally relevant elements, and maintain clarity. Cultural references and idioms should be adapted to local contexts rather than literally translated.

Prompt Creation Annotators generate entirely new prompts in their native language, focusing on originality, clarity, and diversity. They are instructed to consider cultural relevance, ensure contextual completeness, avoid biases, and vary length, style, and complexity. This task expands the dataset beyond reformulations with genuine native language content. Furthermore, the majority of the contributors of the OASST2 were males, with a median age of 26 (Köpf et al., 2023). Therefore, thanks to the "create an original prompt" option, our contributors are able to include topics and tasks that are possibly not represented in the OASST2 dataset already. We include the full annotator guidelines for both tasks in the Appendix A.

3.2. Platform Implementation and Contributor Management

We implement a web-based annotation platform (Figure 2)[1] to facilitate the two complementary tasks described above and evaluation as well, which we do not elaborate on in this work. Key features include task-based workflows, multilingual interface, gamification elements, and separation of prompt evaluation from response generation, which enables independent quality iteration.

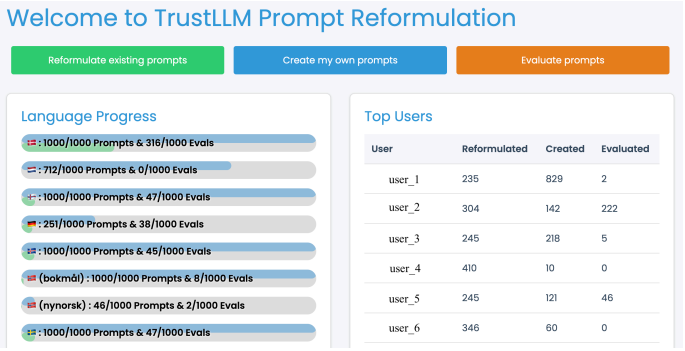

Figure 2: Web-based platform for collecting prompts. The interface provides three main workflows: Reformulate (adapting English prompts), Create (generating original prompts), and Evaluate (not discussed in this paper), with annotation guidelines, real-time progress tracking, multilingual support, and leaderboards to encourage engagement. Usernames are anonymized.

Contributor Recruitment Recent research highlights challenges LLMs pose for crowdsourcing data quality, as contributors may use LLMs to generate responses, introducing model biases and reducing data authenticity (Zhang et al., 2025). Therefore, rather than open crowdsourcing, we recruited annotators from the TrustLLM project[2] and hired students. We conducted online annotation seminars to ensure consistent understanding of reformulation guidelines. This controlled model enabled quality control and appropriate handling of potentially sensitive prompts from OASST2.

3.3. Task and Topic Taxonomy Design

To systematically categorize prompt diversity, we develop a two-dimensional classification system

[1]The platform is open-source and available at https://github.com/AnnikaSimonsen/prompt-reformulation-TrustLLM. Built with Flask (Python), TailwindCSS, DaisyUI, htmx, and Alpine.js, deployed on Heroku with PostgreSQL. The modular design allows easy adaptation for other language pairs or annotation tasks.

[2]https://trustllm.eu/

grounded in established frameworks including MT-Bench (Zheng et al., 2023), ChatGPT usage patterns (Chatterji et al., 2025), and Dolly (Conover et al., 2023) (see Section 2.3 for details).

Our task taxonomy comprises 8 primary categories with 29 subcategories:

1. **Information & Explanation**: 1.1 Factual questions, 1.2 Explanations & tutorials, 1.3 How-to & recipes
2. **Writing**: 2.1 Creative, 2.2 Professional, 2.3 Content generation
3. **Editing & Rewriting**: 3.1 Editing, 3.2 Rewriting, 3.3 Summarization, 3.4 Translation
4. **Reasoning**: 4.1 Logical, 4.2 Critical thinking, 4.3 Common sense
5. **Math**: 5.1 Arithmetic, 5.2 Word problems, 5.3 Advanced mathematics
6. **Coding**: 6.1 Generation, 6.2 Debugging, 6.3 Explanation, 6.4 Algorithm design
7. **Extraction & Classification**: 7.1 Information extraction, 7.2 Classification, 7.3 Sentiment
8. **Planning & Advice**: 8.1 Brainstorming, 8.2 Roleplay, 8.3 Planning, 8.4 Practical guidance, 8.5 Professional advice, 8.6 Personal advice

Key design decisions: First, we merge *Information-seeking* and *Educational explanation* into *Information & Explanation* (1) after pilot testing revealed inter-annotator disagreement, organizing subcategories by answer structure (factual, conceptual, procedural). Second, we explicitly include *How-to & Recipes* (1.3) as a prominent task type. Third, following MT-Bench (Zheng et al., 2023), we elevate Math and Coding to primary categories as distinct reasoning modalities.

Topic Taxonomy We develop a 10-domain taxonomy separating content domains from task types, enabling orthogonal classification. Unlike MT-Bench, which included STEM and Humanities as task categories, we treat these as topic domains: STEM & Technology, Business & Professional, Humanities & Arts, Social Sciences, Education & Learning, Health & Wellness, Lifestyle & Practical, Entertainment & Media, Language & Communication, and General Knowledge.

Classification Methodology We employ multi-label classification with one primary and up to two secondary categories per dimension. Classification uses two LLMs (Gemini 2.5 Pro and GPT-5) as automated annotators, following the "LLM-as-judge" paradigm (Gu et al., 2025), outputting structured JSON with category codes, confidence scores, and reasoning. This dual-LLM approach serves two purposes: comparing classification consistency between different model architectures and identifying ambiguities in our taxonomy design. Initial validation on Faroese and Icelandic samples revealed that Gemini 2.5 Pro produced more accurate classifications for these low-resource languages. Based on this finding, we select Gemini 2.5 Pro for classifying the remaining languages (German, Dutch, Swedish, and Norwegian Bokmål).

Our use of proprietary LLMs (Gemini 2.5 Pro and GPT-5) for classification rather than data generation mitigates data leakage concerns. Since these models classify existing human-written prompts rather than generate training data, potential contamination from their training sets does not affect our dataset's integrity. However, we acknowledge uncertainty about these models' training data composition, which may influence classification performance across languages. To validate both the diversity of our collected prompts and the reliability of our classification approach, we conduct two complementary analyses: Shannon entropy measurements quantifying task and topic coverage, and inter-annotator agreement assessment between Gemini 2.5 Pro and GPT-5.

3.4. Human vs. Synthetic Prompt Quality Evaluation

To assess the quality advantage of human reformulation over purely synthetic approaches, we conduct a blind evaluation comparing human-reformulated prompts with synthetically generated prompts (See Appendix B for the annotator guidelines). We employ the Magpie method (Xu et al., 2024) using Llama 3.3 70B instruct, which generates instruction prompts by prompting aligned LLMs with nothing, allowing the model to autonomously produce diverse instructions without seed examples. For each of three languages (Danish, Faroese, Icelandic), we sample 100 human-reformulated prompts and generated 100 synthetic Magpie prompts, yielding 600 prompts total.

Quality Rating Scale Native speakers evaluate prompts into four categories assessing naturalness and comprehensibility: (1) Natural and comprehensible: sounds like a native speaker wrote it, meaning is clear; (2) Somewhat natural but unclear: grammar and phrasing seem okay, but meaning is confusing or ambiguous; (3) Unnatu-

ral but comprehensible: meaning is understandable, but phrasing sounds "off" (awkward phrasing, literal translations, non-native patterns, mixed-language elements); (4) Incomprehensible: cannot understand the intended meaning (gibberish, wrong language, nonsensical). Evaluators are blind to prompt source (human vs. synthetic) and assess them in randomized order to prevent bias.

To determine whether the distributions of quality ratings differed significantly between human-reformulated and synthetic prompts, we employ the chi-square test of independence. This test assesses whether the overall distribution of ratings across the four quality categories differs by prompt source (human vs. synthetic), evaluating whether the frequency distribution of ratings is independent of the prompt generation method. Statistical significance ($p < 0.05$) indicates that the distributions differ beyond what would be expected by chance. We select the chi-square test because our rating scale, while having numbered categories, is categorical rather than truly ordinal, since categories 2 (natural but unclear) and 3 (unnatural but comprehensible) represent different dimensions of quality that cannot be definitively ordered.

4. Data Collection and Results

Our data collection yielded 6,950 prompts across seven Germanic languages. We report on language coverage and participation patterns, assess the reliability of our classification approach through inter-annotator agreement, analyze task and topic diversity, and evaluate prompt quality through blind comparison with synthetic alternatives.

4.1. Language Coverage and Participation

Table 2 summarizes our language coverage. The 54 contributors produced 6,950 contributions across seven Germanic languages. Five languages achieved 1,000+ prompts, with substantial Dutch coverage (700) and moderate German coverage (250). The dataset contains 5,950 human-created prompts: 2,791 (47%) original and 3,159 (53%) reformulated, plus 1,000 machine-translated Norwegian Nynorsk prompts, with substantial variation across languages reflecting contributor preferences.

Our Norwegian Nynorsk prompts were created by translating the Norwegian bokmål prompts. The Norwegian written variants Bokmål and Nynorsk are very similar, and the Constraint Grammar-based Apertium[3] translation model has achieved state-of-the-art performance up until recently. Despite that better results have been published, e.g., for the encoder-decoder model NorT5 (Samuel et al., 2023), similar rule-based systems are still used commercially. Apertium as also been used to create synthetic data for smaller languages in Spain (Sant et al., 2024). Reasonable performance combined with robust open source licenses warranted the choice.

The machine translated Norwegian Nynorsk prompts were proofread by a native speaker. The translations by Apertium model `nob-nno_e` broadly contained three types of errors. First, the translations into Nynorsk struggled with interrogative pronouns (especially *hvilke(t)*, where the resulting translations were entirely incorrect, or odd. Next, the translations often failed with anaphoric personal pronouns, which must adhere to gender agreement in Nynorsk, but not in Bokmål. Finally, some English names were misunderstood and translated (e.g., "stream deck" became "strøym deck"). Otherwise, mistakes were minor, as some odd, but grammatical, word orders. In the final revision, some word choices were replaced (Nynorsk permissively allows multiple norms). The manual proofreading also uncovered some minor mistakes in the Bokmål originals that were simultaneously fixed.

Language	Samples	Contributors
Swedish	1,000	7 (12)
Norwegian (Bokmål)	1,000	2 (2)
Norwegian (Nynorsk)*	1,000	2 (2)
Danish	1,000	5 (6)
Icelandic	1,000	4 (11)
Faroese	1,000	2 (4)
Dutch	700	6 (10)
German	250	5 (9)
Total	**6,950**	**31 (54)**

Table 2: Language coverage and contributor participation. Contributors column shows male (total) counts. Total unique contributors may be lower due to multiple accounts per individual. *Norwegian (Nynorsk) samples are machine translated from Norwegian (Bokmål) and proofread by a native speaker.

4.2. Inter-annotator Agreement

We evaluated inter-annotator agreement between Gemini 2.5 Pro and GPT-5 on task and topic annotations for 2,000 prompts across Danish and Faroese (Table 3). Agreement was substantial to almost perfect: Cohen's κ = 0.770 for task subcategories, κ = 0.848 for main task categories, and κ = 0.835 for topics. When accounting for partial

[3] http://www.apertium.org

credit (same main category but different subcategory), weighted agreement reached 0.849. Agreement was slightly higher for Danish (86.4% tasks, 87.2% topics) than Faroese (83.4% tasks, 84.5% topics), likely reflecting differences in model training data availability.

Measure	All	DA	FO
Primary Task			
Wtd. agr.	0.85	0.86	0.83
κ (sub.)	0.77	0.79	0.75
κ (main)	0.85	0.87	0.82
Primary Topic			
Agr.	0.86	0.87	0.85
κ	0.84	0.85	0.83
n	2,000	1,000	1,000

Table 3: Inter-annotator agreement metrics for task and topic classification. All Cohen's κ values indicate substantial ($\kappa > 0.61$) to almost perfect ($\kappa > 0.81$) agreement. Abbreviations: Wtd. agr. = Weighted agreement; κ (sub.) = Cohen's κ at subcategory level; κ (main) = Cohen's κ at main category level; Agr. = Agreement score; DA = Danish; FO = Faroese.

The primary source of disagreement involved distinguishing between information-seeking requests (Category 1) and advice-seeking requests (Category 8), which accounted for 45 of 135 main category disagreements. This reflects a known challenge in instruction classification, where queries like "How do I approach X?" can reasonably be interpreted as either seeking factual information or strategic guidance.

4.3. Task and Topic Distribution

Using the task and topic taxonomy framework described in Section 3.3, we classified all prompts (N=6,965) using Gemini 2.5 Pro. The analysis reveals consistent diversity patterns across all seven Germanic languages, with some variations reflecting both annotator expertise and language community characteristics.

To quantify dataset diversity, we calculated Shannon entropy ($H = -\sum_i p_i \log p_i$) for both categorical distributions and token-level distributions. *Categorical distribution entropy*, which measures how evenly prompts are distributed across task and topic categories, reveals high diversity across all languages. Task entropy ranges from 1.87 nats (Norwegian Bokmål) to 2.53 nats (Faroese), with most languages exceeding 2.2 nats, indicating well-distributed task coverage (See Figure 3). Topic entropy ranges from 1.89 nats (German) to 2.14 nats (Faroese) (See Figure 4). These values approach theoretical maxima ($H'_{\text{max}} = 3.37$ for 29 tasks; $H'_{\text{max}} = 2.30$ for 10 topics), confirming that reformulation and creation created diverse coverage across linguistic communities. We additionally analyzed *linguistic entropy* at the token level ($H = -\sum_w p(w) \log_2 p(w)$) to assess vocabulary richness. Token entropy ranges from 8.98 bits (Dutch) to 9.78 bits (Faroese), with vocabulary sizes spanning 1,615 unique tokens (German, $n = 253$) to 6,558 tokens (Norwegian Bokmål, $n = 1{,}001$). Type-token ratios range from 0.194 (Norwegian Nynorsk) to 0.414 (German), with the higher German ratio reflecting the smaller sample size rather than greater lexical diversity. These linguistic entropy values confirm that reformulation and creation produces not only diverse task coverage but also linguistically rich prompts with varied vocabulary, avoiding the repetitive phrasing often observed in purely synthetic datasets.

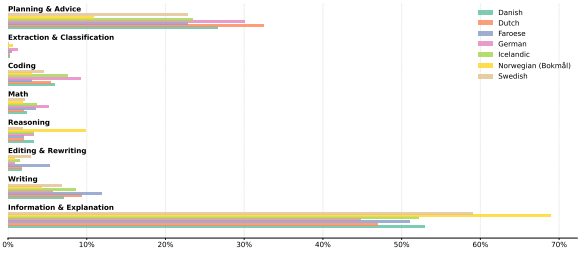

Figure 3: Primary task distribution across seven Germanic languages: Danish (N=1,000), Dutch (N=700), Faroese (N=1,000), German (N=250), Icelandic (N=1,000), Norwegian (Bokmål) (N=1,000) and Swedish (N=1,000).

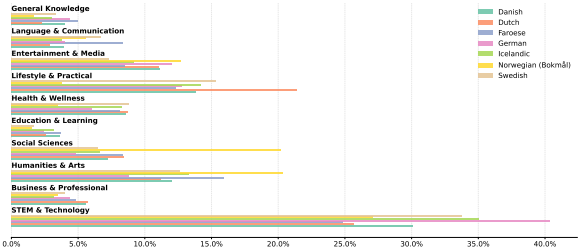

Figure 4: Primary topic distribution across seven Germanic languages: Danish (N=1,000), Dutch (N=700), Faroese (N=1,000), German (N=250), Icelandic (N=1,000), Norwegian (Bokmål) (N=1,000) and Swedish (N=1,000) using Gemini-2.5-Pro.

4.4. Human vs. Synthetic Prompt Quality Evaluation

Using the evaluation methodology described in Section 3.4, native speakers blindly rated 600 prompts (100 human and 100 synthetic per lan-

guage for Danish, Faroese, and Icelandic) on a 4-point naturalness and comprehensibility scale.

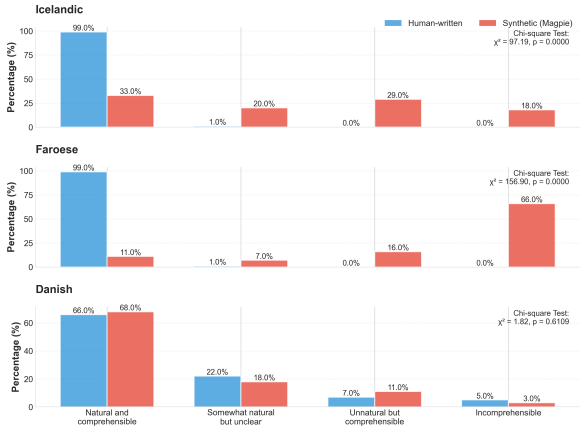

Figure 5: Quality evaluation comparing human-reformulated and synthetic prompts for Danish, Icelandic, and Faroese. Bar charts show the distribution of quality ratings (1=Natural and comprehensible, 2=Somewhat natural but unclear, 3=Unnatural but comprehensible, 4=Incomprehensible) for both human-written and synthetic (Magpie) prompts. Chi-square test statistics are shown for each language.

Results reveal substantial quality differences between human-reformulated and synthetic prompts, with magnitude varying by language and resource availability (Figure 5). For Icelandic and Faroese (the two lowest-resource languages in our dataset), human reformulation demonstrated better naturalness and comprehensibility. Specifically, 99% of human-written prompts in both Icelandic and Faroese received the highest rating (natural and comprehensible), compared to only 33% for Icelandic synthetic prompts and 11% for Faroese synthetic prompts. Chi-square tests confirm these differences are highly significant (Icelandic: χ^2 = 97.19, p < 0.0001; Faroese: χ^2 = 156.90, p < 0.0001).

Danish results showed less pronounced differences, with 66% of human prompts and 68% of synthetic prompts rated as natural and comprehensible (χ^2 = 1.82, p = 0.61). However, post-evaluation discussion revealed that our evaluation guidelines were insufficiently clear. The Danish annotator applied prescriptive linguistic judgments regarding word order and lexical choices, which, in consultation with other native speakers, were confirmed as acceptable in authentic user-LLM interactions. This reflects a limitation in our guideline design: we failed to adequately specify that evaluations should capture descriptive judgments of what real users might naturally write rather than prescriptive ideals. Future evaluations should explicitly instruct annotators to assess whether prompts reflect how real users would naturally interact with an LLM, rather than applying prescriptive grammatical standards.

These findings demonstrate that human reformulation provides critical quality advantages for truly low-resource languages like Faroese and Icelandic, where synthetic generation struggles to produce natural, comprehensible prompts.

5. Discussion

Quality and Diversity Through Reformulation Our reformulation methodology successfully produced high-quality, diverse prompts across seven Germanic languages. Having native speakers create prompts for their own languages fostered a sense of community ownership (Nekoto et al., 2020). This approach also provides a replicable template for other language families, supported by an open-source platform that can be adapted to diverse linguistic contexts.

Our dual-task approach, that offers both reformulation and original prompt creation, accommodated diverse contributor preferences while maintaining dataset diversity. Contributors naturally gravitated toward different tasks: some preferred the creative freedom of generating entirely original prompts, while others found reformulation more engaging as it provided structured starting points. Importantly, many contributors alternated between tasks, with the reformulation task often sparking ideas that they subsequently developed into original prompts. This cross-pollination between tasks proved essential for both contributor engagement and maintaining diversity. The reformulation task ensured broad coverage of established task types from OASST2, while the original creation task enabled contributors to address topics and perspectives absent from the English-language source material.

Small Team Reformulation and Community Engagement A key advantage of our reformulation approach is its ability to preserve dataset diversity, even with smaller contributor teams. While the OASST2 dataset seems to consist largely of prompts within STEM and technology topics, and therefore ours do as well, our small teams (ranging from 2-7 contributors per language) were able to maintain broad task diversity while enhancing topic representation. Our taxonomy analysis revealed that OASST2 exhibits significant STEM bias (30-40% of prompts), and therefore our dataset inherits this distribution.

Our results demonstrate that strategically composed small teams can actually enhance diversity rather than constrain it. For example, in Faroese, a language with limited NLP resources, just four con-

tributors (each with linguistic expertise) achieved exceptional topic diversity (Shannon H' = 2.14) and strong representation in the Humanities & Arts (16%). Contributors' linguistic expertise enriched the dataset, compensating for scarce NLP resources in this language.

This variation in topic distribution, such as Norwegian's high representation of Social Sciences (20%) and Faroese's emphasis on Humanities (16%), illustrates the natural diversity that can emerge from different cultural contexts. By incorporating such variations, our dataset captures a broader spectrum of human knowledge and values.

Crucially, this diversity was achieved *through* reformulation and creation. By enabling annotators to adapt prompts to their cultural and linguistic contexts, reformulation introduces new perspectives while maintaining the diversity of task types. This is particularly beneficial for low-resource languages, where large annotator pools are often not possible. Strategic recruitment of small, diverse teams, paired with reformulation and creation, offers a viable path for creating instruction datasets that represent a wide array of human knowledge and relevant cultural context.

Our findings reveal a counterintuitive pattern: smaller language communities reached the 1,000-prompt target more quickly than larger ones (see Table 2). Communities speaking lower-resource languages, like Faroese, demonstrated higher volunteer engagement, likely reflecting a greater sense of urgency around language preservation and representation in AI systems. In contrast, larger languages required more structured recruitment approaches, including focused annotation seminars, to achieve comparable participation levels. For German, engagement remained limited; we stopped collecting after 250 prompts given that substantial German instruction datasets already exist.

While our methodology emphasizes reformulation over direct translation, practical constraints sometimes necessitate adaptations. For Norwegian Nynorsk, where we had limited contributors, we resorted to translating the Bokmål prompts and proofreading them with the help of a native speaker. This experience teaches us an important lesson: when annotator availability is limited, careful translation with native speaker review can be a preferable alternative to lacking data, even if reformulation remains the ideal approach. We acknowledge that this approach represents a pragmatic departure from our core methodology. However, the linguistic proximity of Bokmål and Nynorsk, which differ primarily in orthographic conventions and some lexical choices, makes translation between them fundamentally different from translating from English.

Impact of Source Distribution Bias Our dataset inherits OASST2's concentration in STEM and technology topics (30–40% of prompts) and Information & Explanation tasks (50–60%). This imbalance may affect downstream model performance: models fine-tuned on our data may perform better on information-seeking tasks than on underrepresented categories like coding, math, or extraction. Future work should explore stratified sampling or targeted prompt creation to rebalance these distributions before fine-tuning. We note that the original prompt creation task partially mitigates this bias, as contributors introduced topics absent from OASST2, but systematic rebalancing remains necessary for training general-purpose models.

Challenges with Community Evaluation While prompt reformulation and creation succeeded, our evaluation component where contributors assessed each other's work faced insufficient participation. Lack of compensation was a key factor; contributors engaged readily in creative reformulation but found the tedious evaluation task less motivating without financial incentives.

Taxonomy Validation and Classification Insights Our systematic classification reveals task and topic distribution patterns in our dataset. Since our prompts were reformulated from OASST2, these patterns may reflect characteristics of the source dataset, though we cannot make definitive claims about OASST2's distribution without direct measurement. Substantial inter-annotator agreement between our two LLM classifiers (Cohen's kappa = 0.769–0.845) validates our taxonomy's robustness. However, disagreements offer insights into instruction classification ambiguities. For example, the Danish prompt *Hvem er den bedste spiller i den danske Superliga?* ("Who is the best player in the Danish Superliga") was categorized by Gemini 2.5 Pro as factual but by GPT-5 as critical thinking, reflecting genuine ambiguity about whether "best player" requires factual metrics or subjective judgment. Such disagreements show that some prompts are inherently multifaceted and context-dependent.

6. Future Work

We generated synthetic responses using Deepseek R1 (685B) (DeepSeek-AI, 2025), but quality was inadequate for immediate release due to errors in Icelandic and Faroese and hallucinations on culturally-specific questions. Phase

1 releases human-reformulated prompts; Phase 2 will focus on response quality assurance. An important next step is evaluating downstream impact: once validated responses are available, we plan to fine-tune multilingual LLMs on our prompt-response pairs and assess whether human-reformulated training data yields measurable performance improvements over translated or synthetic alternatives. Our prompts enable immediate applications including test sets for multilingual evaluation and examples of authentic native formulations. Future work includes exploring few-shot prompting with human examples (Longpre et al., 2023), retrieval-augmented generation, and hybrid human-in-the-loop refinement (Kim et al., 2023). For researchers adapting our methodology, we recommend automated tagging followed by stratified sampling to ensure balanced task distribution (Gadre et al., 2023).

7. Conclusion

We introduced a prompt reformulation methodology for creating high-quality, diverse datasets for low-resource Germanic languages. Our dual-task approach, that combines reformulation with original prompt creation, enabled contributors to maintain broad task coverage from the source dataset while addressing topics and cultural contexts absent from English-language materials. This approach overcomes the challenge of small annotator pools, allowing broad task coverage even with limited contributors.

Our results show that human reformulation significantly outperforms synthetic methods, especially for low-resource languages like Icelandic and Faroese, where synthetic prompts often fail to capture cultural nuances. For larger languages like Danish, the difference is less prominent. Beyond the datasets themselves, our systematic classification quantifies the task and topic distribution in our reformulated dataset, which may provide insights into distribution patterns in datasets derived from OASST2. Our methodology provides a practical, scalable path for creating culturally appropriate instruction data and contributes to a more inclusive AI ecosystem. The release of our datasets, platform code, and taxonomy framework offers resources for expanding this work to other language families globally, with particular promise for language families with one dominant high-resource language and multiple lower-resource relatives.

8. Acknowledgements

We thank all community contributors who participated in prompt reformulation and creation across seven Germanic languages, in particular the members of the TrustLLM consortium and the student annotators. We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript. AS and MS were supported by the European Commission under grant agreement no. 101135671.

9. Limitations

Several limitations should be acknowledged. First, we release only prompts, not complete prompt-response pairs. The synthetic responses require quality validation before release. Second, our controlled contributor model limits scale and diversity compared to open crowdsourcing. For small contributor pools, individual preferences may disproportionately influence dataset characteristics. Third, prompt quality evaluation was only conducted for three languages (Danish, Faroese, Icelandic). This restriction reflects practical constraints: suitable evaluators required both native-speaker fluency and familiarity with LLM-generated text, a combination we could not secure for Swedish, German, and Dutch within the project timeline. Additionally, our evaluation guideline design was insufficient for the Danish assessment: we failed to clearly specify that evaluators should apply descriptive rather than prescriptive linguistic standards when judging naturalness. Fourth, our dataset exhibits certain distributional patterns, which we systematically quantified. Task distribution shows concentration in Information & Explanation (50-60%), while Coding (5%), Math (3-5%), and Extraction (<2%) have lower representation. Topic distribution shows higher coverage of STEM (30-40%), with lower coverage of Business (5-10%) and Education (2-5%). Since our prompts were reformulated from OASST2, these patterns may reflect the source dataset, though we cannot definitively characterize OASST2 without direct measurement. These distribution patterns may affect downstream model performance, potentially favoring information-seeking over technical tasks like coding or mathematical reasoning.

10. Ethics Statement

This research involved human participants as contributors and evaluators. All contributors were informed about project goals, their role in creating AI training data, and planned public release under open-source license. Participation was voluntary with withdrawal rights. Students who were hired received payment by their institutions; project members volunteered. All contributor information has been anonymized.

The source data for reformulation came from the OpenAssistant Conversations (OASST2) dataset, which contains some prompts addressing sensitive topics or using informal language. We addressed this through controlled recruitment rather than open crowdsourcing, enabling us to provide appropriate context and support to contributors. Our annotation seminars explicitly discussed how to handle potentially sensitive content, and contributors were encouraged to skip or modify prompts they found inappropriate.

Our dataset reflects biases from source data and contributors. Mitigation factors include preserving OASST2's task diversity, achieving gender balance (46% female), and enabling quality oversight through controlled recruitment. Nevertheless, users should recognize it represents a specific contributor group and may not capture full language diversity.

We release the prompt datasets under an open-source license to maximize research accessibility and enable scrutiny. Users should exercise responsibility, as some prompts may be inappropriate for certain contexts, and researchers must ensure applications align with ethical principles and local norms.

This work contributes to linguistic diversity in AI by providing resources for historically underserved languages. However, meaningful inclusion requires more than data availability; it demands ongoing community involvement, respect for cultural contexts, and recognition that technology development should serve rather than be imposed upon linguistic communities. We do not release synthetic responses due to quality concerns that could lead to harmful model behaviors.

11. Bibliographical References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron Chatterji, Tom Cunningham, and David Deming. 2025. How People Use ChatGPT . Working Paper 34255, National Bureau of Economic Research.

Hao Chen, Abdul Waheed, Yousef Zhang, Zhaozhuo Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. 2024. On the Diversity of Synthetic Data and its Impact on Training Large Language Models.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Iben Nyholm Debess, Alina Karakanta, and Barbara Scalvini. 2025. What's Wrong With This Translation? Simplifying Error Annotation For Crowd Evaluation. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 42–47, Tallinn, Estonia. The University of Tartu Library.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J Ratner, Shuran Song,

Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, pages 27092–27112. Curran Associates, Inc.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam {de Lhoneux}, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura {Cabello Piqueras}, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 6997–7013, United States. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning Large Language Models through Synthetic Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700, Singapore. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. 2024. OctoPack: Instruction Tuning Code Large Language Models.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaL-*

iDa), pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Aleix Sant, Daniel Bardanca, José Ramom Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier Garcia Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Training and Fine-Tuning NMT Models for Low-Resource Languages Using Apertium-Based Synthetic Corpora. In *Proceedings of the Ninth Conference on Machine Translation*, pages 925–933, Miami, Florida, USA. Association for Computational Linguistics.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Natural Questions in Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4488–4496, Marseille, France. European Language Resources Association.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. *GitHub repository*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. OpenChat: Advancing Open-source Language Models with Mixed-Quality Data.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

Yichi Zhang, Jinlong Pang, Zhaowei Zhu, and Yang Liu. 2025. Evaluating LLM-corrupted crowdsourcing data without ground truth.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

12. Language Resource References

Language Resources

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Javier de la Rosa, Vladislav Mikhailov, Lemei Zhang, Freddy Wetjen, David Samuel, Peng Liu, Rolv-Arild Braaten, Petter Mæhlum, Magnus Breder Birkenes, Andrey Kutuzov, Tita Enstad, Hans Christian Farsethås, Svein Arne Brygfjeld, Jon Atle Gulla, Stephan Oepen, Erik Velldal, Wilfred Østgulen, Lilja Øvrelid, and Aslak Sira Myhre. 2025. The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 544–560, Tallinn, Estonia. University of Tartu Library.

Oskar Holmström and Ehsan Doost. 2023. Making Instruction Finetuning Accessible to Non-English Languages: A Case Study on Swedish Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

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, Tórshavn, Faroe Islands. University of Tartu Library.

Kasper Junge, Martin Exner, Martin Sønderlev Christensen, and Danni Dromi. 2024. SkoleGPT-Instruct: A Danish Instruction Dataset. https://huggingface.co/datasets/kobprof/skolegpt-instruct.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith

Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. HuggingFace H4 Stack Exchange Preference Dataset.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025. A Collection of Question Answering Datasets for Norwegian. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 397–407, Tallinn, Estonia. University of Tartu Library.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval. *arXiv preprint arXiv:2104.12741*.

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.

Bjoern Patz. 2024. UltraChat DE: German Translation of UltraChat. https://huggingface.co/datasets/bjoernp/ultrachat_de.

Niels J. Rouws, Svitlana Vakulenko, and Sophia Katrenko. 2022. Dutch SQuAD and Ensemble Learning for Question Answering from Labour Agreements. In *Artificial Intelligence and Machine Learning*, pages 155–169, Cham. Springer International Publishing.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

13. Appendix

A. Annotation Guidelines for Prompt Reformulation and Creation

The following guidelines were provided to annotators on the annotation platform for the two prompt collection tasks.

A.1. Reformulation Guidelines

Welcome to the Reformulation task! Your goal is to take an English prompt and reformulate it in your chosen language. Here are some key guidelines to follow:

1. **Flexibility in Meaning**: You have the freedom to change the meaning of the original prompt if you wish. For example, if the original prompt asks for a chocolate cake recipe, you could reformulate it to ask for a lasagne recipe instead. However, you're also welcome to maintain the original meaning if you prefer. To improve diversity in the data, we encourage you to change the meaning of the original prompt, but try to stay within the same task category.

2. **Natural Language Use**: Your reformulation should sound natural in the target language. Avoid literal translations that might sound awkward or unnatural.

3. **Cultural Context Consideration**: Be mindful of cultural differences. Some concepts or idioms might need to be adapted to make sense

in the target language and culture. Try to make references to local culture if possible.

4. **Tone and Style Flexibility**: You have the option to maintain or change the tone (formal, casual, etc.) and style (descriptive, questioning, etc.) of the original prompt as you see fit.

5. **Clarity is Key**: Make sure your reformulation is clear and unambiguous. If the original has any vague parts, aim to clarify them if possible.

6. **Length Consideration**: It's okay if your reformulation is slightly longer or shorter than the original.

7. **Handle Technical Terms Carefully**: If the prompt contains technical terms, make sure you translate them accurately or use accepted terms in the target language.

8. **Grammar and Errors**: Unintentional grammatical mistakes are allowed to remain in your reformulation. However, do not deliberately insert grammatical errors.

9. **When in Doubt, Refresh the Prompt**: If you're unsure about how to reformulate a particular prompt, it's better to skip it than to submit an inaccurate reformulation.

The goal is to create a prompt in the target language that allows for creativity while still maintaining the spirit of the task.

A.2. Prompt Creation Guidelines

Welcome to the Prompt Creation task! Your goal is to create original prompts in your chosen language. These prompts will be used to train and evaluate AI language models. Here are some key guidelines to follow:

1. **Originality**: Create prompts that are original and unique. Avoid copying existing prompts or well-known quotes.

2. **Clarity**: Write clear and unambiguous prompts. The intent of your prompt should be easily understandable.

3. **Diversity**: Try to create a diverse range of prompts. This could include different topics, styles (questions, statements, scenarios), and complexity levels.

4. **Cultural Relevance**: Consider creating prompts that are relevant to the culture and context of the language you're writing in.

5. **Avoid Bias**: Be mindful of potential biases in your prompts. Aim for neutral language that doesn't unfairly represent or exclude any groups.

6. **Length**: Prompts can vary in length, but try to keep them concise. A good range is typically between 10 to 50 words.

7. **Purpose**: Consider the potential use of the prompt. It could be for generating a response, continuing a story, answering a question, or describing something.

8. **Creativity**: Don't be afraid to be creative! Interesting or thought-provoking prompts can lead to more engaging responses.

9. **Appropriate Content**: Ensure your prompts are appropriate for a general audience. Avoid explicit content, hate speech, or overly controversial topics.

10. **Grammar and Spelling**: Double-check your prompts for correct grammar and spelling in the target language.

11. **Avoid Personal Information**: Don't include any personal or identifying information in your prompts.

12. **Contextual Completeness**: Ensure that your prompt provides enough context to stand alone. It shouldn't rely on external information not included in the prompt itself.

The prompts created will be used to train and test AI models. Contributions are crucial in developing AI systems that can understand and generate text in multiple languages.

B. Prompt Quality Evaluation Guidelines

The following guidelines were provided to native-speaker evaluators for the blind quality assessment of human-reformulated and synthetic prompts. Evaluators were blind to prompt source and assessed prompts in randomized order.

For each prompt, select the category that best describes its quality:

1. **Natural and comprehensible**: Sounds like a native speaker wrote it; the meaning is clear.

2. **Somewhat natural but unclear**: Grammar and phrasing seem okay, but the meaning is confusing or ambiguous.

3. **Unnatural but comprehensible**: You understand what they are asking, but it sounds "off" — there is awkward phrasing, literal translations, non-native patterns, or non-target-language words mixed in.

4. **Incomprehensible**: Cannot understand the intended meaning (e.g., the text is gibberish, in the wrong language, or nonsensical).

Evaluators were instructed to assess prompts based on how natural they would sound as real user queries to a language model, rather than applying prescriptive grammatical standards.