MaritimEmails: A Synthetic Dataset for Maritime Chartering Correspondence

Kevin Bründler, Simon Clematide
University of Zurich
Andreasstrasse 15, 8050 Zürich, Switzerland
kevin.bruendler@uzh.ch, simon.clematide@cl.uzh.ch

Abstract

We introduce MaritimEmails, a large-scale synthetic corpus of 19,817 English-language email threads simulating maritime chartering negotiations between brokers and charterers. Email remains a dominant medium for business communication, yet no public corpora exist for this highly specialized domain due to confidentiality constraints. To address this gap, we generate domain-plausible negotiation exchanges using five contemporary language models under multiple prompting strategies, including Attribute Prompting and Base–Refine (BARE) approaches. Each thread includes structured annotations for vessels, ports, commodities, and Incoterms, enabling supervised training for information extraction and related tasks. Our comparative evaluation covering lexical and semantic diversity, sentiment balance, and verbosity shows that BARE generation increases linguistic variation while maintaining coherence. However, all models exhibit a systematic positivity bias, yielding less negative sentiment than is observed in the Enron reference corpus and likely also in many real negotiation settings. Baseline information extraction experiments with GLiNER and generative Qwen models yield up to 0.86 macro F1 on entity extraction, supporting the dataset's usefulness. MaritimEmails, together with prompts, scripts, and documentation, is released for research use.

Keywords: Synthetic Data, Email Corpora, Maritime Chartering, Information Extraction, Language Resources

1. Introduction

1.1. Motivation

Although email remains a primary medium for professional communication, with more than 300 billion messages sent each day (Statista Research Department, 2024), publicly available email datasets are scarce. Most research still relies on the Enron corpus (Klimt and Yang, 2004), which dates from the early 2000s and only partially reflects current writing styles or specialized business domains. This lack of data limits the development of NLP systems that can process professional communication in specific industries. Similar shortages have motivated research on synthetic data generation as an alternative source of supervised training data across multiple domains (Lu et al., 2023; Long et al., 2024a; Li et al., 2024).

Maritime chartering is one such domain. Around 80% of global trade by volume is transported by sea (United Nations Conference on Trade and Development, 2024), and most commercial arrangements between shipbrokers and charterers take place by email (Zhou et al., 2021; Shin et al., 2018). These exchanges use specialized terminology and abbreviations (for example, *LOA*, *DWT*, *MV*, and *MT*) and follow a concise, negotiation-oriented writing style. Because authentic maritime correspondence is commercially sensitive, no open dataset of such emails is publicly available. As a result, models trained on general text may generalize poorly when applied to this domain.

1.2. Contribution

This paper presents **MaritimEmails**, a synthetic corpus of 19,817 email threads that simulate maritime chartering negotiations in which brokers and charterers exchange proposals and counteroffers on freight rates, vessel specifications, and charterparty conditions. We use several generation strategies, including Attribute Prompting (AttrPrompting) (Yu et al., 2023) and the Base-Refine (BARE) approach (Zhu et al., 2025), with five contemporary language models to produce domain-plausible maritime correspondence. Table 1 summarizes the number of email threads generated per model and method. The dataset includes structured annotations for entities such as vessel names, ports, and cargo types. It is intended to serve a dual purpose: as *training data* for building domain-specific NLP models in a setting where no public corpora exist, and as a *benchmark* for evaluating information extraction methods across varying levels of linguistic complexity. Predefined train/dev/test splits and the systematic variation across generation methods support both uses.

To evaluate the dataset's usefulness, we establish information extraction baselines using both span-based NER models (GLiNER (Zaratiana et al., 2023)) and generative approaches. These experiments show that the generation strategy strongly influences both textual diversity and extraction performance: BARE-generated emails exhibit greater linguistic diversity but are more challenging for entity extraction. Our main contributions are threefold: (1) a large-scale synthetic email dataset for mar-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7556–7567
11–16 May 2026.

Method	Mistral	DeepSeek	Claude	GPT-4	Gemini	Row Total
AttrPrompting	989	998	994	1000	1000	4,981
BARE (Llama-3.2-3B)	978	994	991	998	988	4,949
BARE (Llama-3.1-8B)	980	1000	993	997	994	4,964
Few-Shot	494	498	500	499	485	2,476
Zero-Shot	469	498	500	499	481	2,447
Total per Model	**3,910**	**3,988**	**3,978**	**3,993**	**3,948**	**19,817**

Table 1: Synthetic email thread generation counts by model and methodology.

itime chartering with structured entity annotations; (2) a comparative analysis of generation methods and their effects on diversity and stylistic characteristics; and (3) baseline results demonstrating the dataset's value for information extraction.

Paper Structure Section 2 reviews related work on email corpora and synthetic data generation. Section 3 describes the dataset design, generation process, and annotation scheme. Section 4 presents the evaluation framework and main results, and Section 5 reports the information extraction experiments. Section 6 discusses limitations and outlines directions for future work. All data, prompts, and annotation scripts are released for research use.

2. Related Work

2.1. Genuine Email Datasets

Publicly available email datasets remain scarce despite the importance of email communication in business contexts. The Enron corpus (Klimt and Yang, 2004), released in 2004, continues to serve as one of the main public resources for email-related NLP research. It has been used extensively for tasks such as authorship attribution (Fabien et al., 2020) and entity resolution in email conversations (Dakle and Moldovan, 2020). A more recent effort, the MAILEX dataset (Srivastava et al., 2023), adds structured annotations for event extraction but still relies on Enron data. Related resources in adjacent areas include the EmailSum dataset for abstractive email thread summarization (Zhang et al., 2021), which provides 2,549 annotated threads but focuses on general workplace correspondence rather than domain-specific negotiation. Existing resources therefore do not adequately represent contemporary business communication, especially in specialized domains such as maritime chartering, which has its own terminology and communicative conventions.

2.2. Synthetic Data Generation

Advances in synthetic text generation provide new ways to address data scarcity. This trend has been documented in recent survey work on LLM-driven synthetic data generation, including methods for data creation, curation, and evaluation (Long et al., 2024b; Nadăş et al., 2025). Attribute Prompting (AttrPrompting) (Yu et al., 2023) enables fine-grained control over textual attributes through parameterized prompts, producing diverse outputs while maintaining attribute consistency. A key challenge is crafting prompts that incorporate sufficient domain knowledge while still allowing variation. The Base-Refine (BARE) approach (Zhu et al., 2025) addresses the trade-off between quality and diversity through a two-step generation process: a base model (that is, a model that was not instruction-tuned) first generates diverse drafts, potentially at the cost of inconsistencies, and an instruction-tuned model then refines them for coherence and fluency, while preserving core content. This combination yields text that is more varied than standard instruction-tuned outputs yet more consistent than unrefined base generations. Recent work has also explored hybrid or multi-model generation frameworks to enhance diversity and control (Veselovsky et al., 2025; Li et al., 2024, 2023; Lu et al., 2023).

2.2.1. Synthetic Email Generation

Research on synthetic email generation remains limited. Most existing studies focus on stylistic analysis rather than systematic generation frameworks. For instance, Liu et al. (2022) examined human perceptions of AI-generated emails and found that recipients reacted negatively when they knew that an AI system was involved. Similarly, Li et al. (2025) compared human- and model-written emails and reported that LLM-generated emails tend to be more verbose, more formal, and more stylistically uniform, but less linguistically diverse. Other work has focused mainly on narrow applications such as spam generation for classifier training (Heiding et al., 2024; Karanjai, 2022) or personalized email completion (Kumar et al., 2024). To our knowledge, no study has proposed a framework for generating

domain-specific email corpora. Existing research also suggests that synthetic text generation often reduces linguistic diversity and introduces systematic stylistic biases (Guo et al., 2023, 2024; Chen et al., 2024).

3. Resource Description and Generation Pipeline

Creating synthetic chartering emails requires careful attention to domain plausibility and controlled variation. We designed the corpus according to four main principles. First, the emails should use industry-specific terminology, including vessel prefixes (MV, MT, SS), maritime abbreviations (LOA, DWT, pdpr), Incoterms (standardized trade terms such as FOB, CIF, and DDP that define delivery responsibilities), and the concise style typical of shipping negotiations. Second, each thread should reflect plausible negotiation dynamics through iterative offers and counteroffers, conditional clauses, and a logical progression from initial inquiry to fixture or failure. Third, maritime entities should remain consistent within a thread while varying across the dataset: vessels retain their size specifications, port pairs represent plausible trade routes, and freight rates correlate with voyage distance. Fourth, to reduce the risk of overfitting to narrow stylistic patterns, we introduced variation in writing style, formality, and English proficiency, including minor typos and grammatical variation to better reflect real-world business communication. Similar preprocessing and classification techniques have been explored on proprietary shipbrokers' email corpora (Papageorgiou et al., 2024), underscoring the practical relevance of realistic message structure.

To improve domain plausibility, we grounded the generation process in genuine maritime data sources. Vessel specifications were sampled from the Global Cargo Ships Dataset (Ibrahim, 2023), providing ship names, dimensions, and cargo capacities. Port locations were drawn from the Shipping Ports Around The World dataset (Naik, 2023), which contains over 450 major ports worldwide; we geocoded these ports and used the Haversine formula to obtain approximate inter-port distances as a coarse proxy for voyage length and freight-rate estimation. Sender and recipient identities combine the 250 most common male and female first names with 400 frequent US surnames from the Name Dataset (Remy, 2021), enabling over 100,000 unique identity pairings. We selected US surnames because they were readily available in structured form. However, this choice limits the cultural diversity of sender identities and may affect entity extraction performance, as models could learn to associate name patterns with specific entity types. Future iterations could incorporate more internationally diverse name lists to better reflect the global nature of maritime trade. These structured datasets, combined with domain knowledge from scraped maritime news articles and Wikipedia content, provided the foundation for generating emails that balance domain plausibility with the controlled variation needed for effective model training.

To investigate how different language models affect the characteristics of synthetic maritime emails, we use five contemporary models: Mistral (Jiang et al., 2024), DeepSeek (Liu et al., 2024), Claude (Anthropic, 2024), GPT-4 (Achiam et al., 2023), and Gemini (Team et al., 2023). Applying identical generation strategies across all models allows us to compare model-specific effects on textual diversity, sentiment distribution, and linguistic patterns. Model scaling has well-documented effects on both language diversity and stability (Kaplan et al., 2020; Shumailov et al., 2024), which motivates evaluating multiple model sizes in our setup. For the BARE methodology, we use Llama-3.1-8B and Llama-3.2-3B (Grattafiori et al., 2024) as base generators, taking advantage of their capacity for varied outputs before refinement by the instruction-tuned models listed above.

Data Format Each record is represented as a JSON object containing an `email_chain` array and a `labels` object with structured entity annotations. Figure 1 shows a truncated example; the full schema and a complete seven-email thread are documented in the repository.[1]

4. Dataset Analysis and Evaluation

MaritimEmails contains threads of varying length that simulate negotiations and inquiries between shipbrokers and charterers. Threads contain between 1 and 35 emails, with an average of 5.2 messages per thread, totaling 103,705 individual emails across the corpus.

Generation Distribution Table 1 shows that the dataset is balanced across both models and generation methods. Controlled generation approaches (AttrPrompting and BARE variants) account for about 75% of the corpus, with nearly 15,000 threads, while zero-shot and few-shot methods contribute the remaining 25%. Each model generated between 3,910 and 3,993 email chains, so that no single model disproportionately shapes the dataset.

[1] https://github.com/ZurichNLP/MaritimEmails

```
{"email_chain": [
  {"from": "m.mason@sealinetrading.com",
   "to": "n.rosas@globalmaritime.com",
   "subject": "Sugar Cargo -
Cadiz to Heiligenhafen",
   "timestamp": "2014-08-13 09:23",
   "body": "Hi Nate,\n\nLooking for a vessel to
     carry 12,277MT of sugar from Cadiz to
     Heiligenhafen. Need loading window around
     end of August.\n\nBest regards,\nMia"},
  {"from": "n.rosas@globalmaritime.com",
   "to": "m.mason@sealinetrading.com",
   "subject": "Re: Sugar Cargo - Cadiz to ...",
   "timestamp": "2014-08-13 10:45",
   "body": "Dear Mia,\n\nCan offer MV GEMMA for
     yr cargo.\n- Rate: EUR44 PMT CIF\n- Laycan:
     25-30 August\n- Demurrage: EUR 15,000 pd
     pro rata\n\nBest rgds,\nNate"},
  ... ],
 "labels": {
  "commodity": "Sugar",
  "load_port": "Cadiz",
  "discharge_port": "Heiligenhafen",
  "cargo_size": "12277MT",
  "incoterm": "CIF",
  "vessel": "GEMMA", "dwt": "313049",
  "final_freight_quote": "42.5",
  "final_freight_quote_currency": "EUR",
  "laytime_start_date": "2014-08-25",
  "laytime_end_date": "2014-08-31",
  "demurrage": "15000",
  "demurrage_currency": "EUR"}}
```


Figure 1: Truncated example record showing the first two emails of a negotiation thread and the associated structured entity labels.

4.1. Intrinsic Evaluation

We evaluate the dataset using several complementary measures that capture both surface-level linguistic properties and broader semantic patterns. Our analysis focuses on verbosity, textual diversity, and sentiment distribution.

4.1.1. Verbosity Metrics

We quantify structural characteristics using NLTK tokenization (Bird et al., 2009) to compute the average number of words per email, average sentences per message, and average words per sentence. These metrics provide a rough proxy for communication efficiency and stylistic variation across generation approaches and model configurations.

The generated emails vary in length and complexity across models and generation methods. Table 2 reports average words per email, sentences per email, and words per sentence for all configurations. Zero-shot prompting yields the widest range of outputs, with Claude producing the shortest emails (38 words on average) and DeepSeek the longest (86 words). In contrast, BARE methods appear to have a normalizing effect, consistently producing shorter and more uniform messages across models. This brevity is broadly consistent with business communication, which often favors information-dense writing. At the sentence level, Claude generates the fewest sentences per email, while Gemini and GPT-4 produce longer sentences on average. These differences also have practical implications for deployment: longer outputs from models such as DeepSeek and GPT-4 increase token usage, which affects API cost and processing time when generating large-scale synthetic datasets.

4.1.2. Lexical Diversity

We measure vocabulary variety using the distinct-n metrics introduced by Li et al. (2015), which calculate the ratio of unique n-grams to total n-grams. These metrics are widely used for short generated texts such as emails because they partially control for length while capturing repetition. We report distinct-1 (unique unigrams, measuring word-level diversity), distinct-2 (unique bigrams, measuring phrasal variety), and distinct-3 (unique trigrams, assessing longer expressions). Together, these metrics capture lexical repetition, a known limitation of neural text generation systems (Chen et al., 2024; Padmakumar and He, 2023). Related studies have also proposed standardized diversity benchmarks and normalization schemes for comparing LLM outputs (Guo et al., 2024, 2023; Shaib et al., 2024).

As shown in Table 3, lexical diversity varies substantially across generation methods and models. Across all evaluated language models, BARE methods yield higher distinct-n scores than other prompting strategies, with clear gains over AttrPrompting. Using Llama-8B as the base generator, BARE achieves an average improvement of 77.5% across distinct-n metrics (85.4% for distinct-1, 80.9% for distinct-2, and 66.2% for distinct-3). The smaller Llama-3B base model produces similar gains, averaging 69.5%. Notably, BARE (Llama-8B) refined with Claude attains phrasal variation comparable to that of the Enron reference corpus.

Clear model-specific patterns also appear across generation methods. Claude shows the highest lexical diversity under the more structured approaches (AttrPrompting and BARE), whereas GPT-4 maintains relatively high distinct-n scores under minimal prompting. The comparison between structured and minimal prompting therefore suggests model-specific dependencies. DeepSeek exhibits a sharp reduction in diversity, with its distinct-3 score falling from 0.708 (BARE Llama-8B) to 0.121 (few-shot) and 0.065 (zero-shot), corresponding to decreases of 83% and 91%, respectively. Gemini shows similar but less pronounced declines. These results suggest that some models depend more strongly on explicit structural guidance to produce diverse text, while others retain moderate diversity across prompting strategies.

Method	Model	Words/Email	Sentences/Email	Words/Sentence
AttrPrompting	Claude	50.58	4.12	12.27
	DeepSeek	69.93	**5.74**	12.17
	Gemini	49.77	4.35	11.45
	GPT-4	**77.33**	5.31	**14.57**
	Mistral	55.81	4.96	11.26
BARE (Llama-3B)	Claude	45.66	3.58	12.76
	DeepSeek	48.18	3.76	12.80
	Gemini	**51.39**	**3.84**	**13.37**
	GPT-4	49.47	3.80	13.01
	Mistral	46.73	3.67	12.73
BARE (Llama-8B)	Claude	43.50	3.30	13.18
	DeepSeek	46.03	3.52	13.07
	Gemini	**48.87**	3.56	**13.73**
	GPT-4	47.37	**3.60**	13.16
	Mistral	44.73	3.42	13.09
Few-shot	Claude	60.94	3.83	**15.92**
	DeepSeek	**73.85**	**5.74**	12.87
	Gemini	61.74	4.71	13.11
	GPT-4	70.46	5.50	12.81
	Mistral	64.34	5.10	12.61
Zero-shot	Claude	38.08	3.68	10.34
	DeepSeek	**86.17**	**7.01**	12.30
	Gemini	54.07	5.21	10.38
	GPT-4	75.17	6.04	**12.45**
	Mistral	61.50	5.54	11.10

Table 2: Comparison of verbosity metrics (average words per email, sentences per email, and words per sentence). Bold values indicate the highest value within each generation method. The results show substantial variation in output length, with zero-shot prompting producing the widest range (38.08–86.17 words per email), while BARE methods yield more similar values across models.

4.1.3. Semantic Diversity

Beyond lexical variety, we assess conceptual breadth using semantic embeddings. We use `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) to generate vector representations of the first email in each thread, which typically establishes the thematic scope of the conversation. Semantic diversity is defined as one minus the average pairwise cosine similarity among these embeddings, where higher values indicate broader topical dispersion.

The analysis of semantic diversity reveals clear trends. BARE methods achieve 36–37% higher semantic diversity than AttrPrompting and exceed the Enron reference value by approximately 40% (0.308 vs. 0.220). This suggests that BARE-generated emails cover a broader conceptual range than business correspondence in the Enron corpus. Although this wider range may reduce domain specificity, it may also benefit the training of more general-purpose information extraction models. A comparison of human and synthetic unigram diversity shows that even the best configuration—BARE Llama-8B refined with Claude—reaches only 85.6% of the Enron unigram diversity score (0.083 vs. 0.097). This limitation appears inherent to current language models, which recombine common words into varied phrases but still draw on a narrower vocabulary than human writers.

The following sections examine the emotional tone of the generated emails, complementing the lexical and semantic analysis presented above.

4.1.4. Sentiment Analysis

To examine emotional tone and its variation across generation methods, we use an ensemble approach that combines Flair (Akbik et al., 2019) and RoBERTa-large (Hartmann et al., 2023) with equal weights. This dual-model strategy reduces model-specific biases and provides robust sentiment classification into negative, neutral, and positive categories. For consistency, sentiment is analyzed at the thread level by concatenating all messages, thereby capturing the overall conversational tone rather than message-level sentiment in isolation.

Sentiment and Positivity Bias The sentiment results reveal a key limitation of synthetic maritime emails relative to the Enron reference corpus: a pervasive positivity bias. Across all generation methods and model configurations, positive sentiment

Method	Model	Distinct-1	Distinct-2	Distinct-3	Semantic Div.
Human	Enron	0.097	0.526	0.778	0.220
AttrPrompting	Claude	**0.059**	**0.371**	**0.624**	**0.243**
	DeepSeek	0.033	0.200	0.385	0.200
	Gemini	0.041	0.240	0.392	0.238
	GPT-4	0.039	0.268	0.524	0.204
	Mistral	0.033	0.186	0.355	0.236
BARE (Llama-3B)	Claude	**0.075**	**0.454**	**0.760**	0.307
	DeepSeek	0.066	0.397	0.690	**0.309**
	Gemini	0.064	0.408	0.700	0.301
	GPT-4	0.073	0.430	0.735	0.303
	Mistral	0.064	0.393	0.692	0.306
BARE (Llama-8B)	Claude	**0.083**	**0.485**	**0.783**	**0.308**
	DeepSeek	0.070	0.414	0.708	0.304
	Gemini	0.070	0.423	0.703	0.295
	GPT-4	0.077	0.450	0.749	0.302
	Mistral	0.068	0.414	0.708	**0.308**
Few-shot	Claude	0.019	0.133	0.286	0.203
	DeepSeek	0.008	0.053	0.121	0.161
	Gemini	0.017	0.099	0.192	**0.233**
	GPT-4	**0.026**	**0.177**	**0.386**	0.210
	Mistral	0.015	0.093	0.202	0.220
Zero-shot	Claude	**0.017**	0.117	0.234	**0.201**
	DeepSeek	0.005	0.030	0.065	0.022
	Gemini	0.011	0.073	0.143	0.127
	GPT-4	0.015	**0.123**	**0.285**	0.138
	Mistral	0.008	0.058	0.137	0.111

Table 3: Textual diversity metrics benchmarked against the Enron corpus. Distinct-n metrics measure lexical diversity as the ratio of unique n-grams to total n-grams, where higher values indicate greater vocabulary variation. Semantic diversity captures conceptual variety using sentence embeddings. BARE methods use either Llama-3B or Llama-8B as base models for initial generation, followed by refinement. Bold values indicate the highest score within each generation method. The Enron corpus serves as a human-written reference point for general business email. Results show that BARE approaches consistently achieve higher diversity scores than conventional prompting methods, with BARE (Llama-8B) approaching the Enron reference in multi-word expression diversity.

dominates the corpus, accounting for 50–80% of generated content, while negative sentiment rarely exceeds 10%. This distribution contrasts with the Enron corpus, which shows a more balanced sentiment profile with 44.6% positive, 27.1% neutral, and notably, 28.3% negative sentiment.

The extent of this positivity bias varies across models and becomes stronger during BARE refinement. Claude produces the most balanced sentiment distribution across generation methods. In contrast, GPT-4 shows a pronounced positivity bias and rarely generates negative sentiment. During refinement, this bias intensifies; despite explicit instructions to preserve tone, refinement models often convert negative messages into neutral or positive ones. As shown in Figure 2, GPT-4 as a refiner shifts 69% of initially negative emails to either neutral or positive sentiment, increasing the overall share of positive emails from 58% to 76%. Notably, the refinement step does not merely neutralize negative sentiment but appears to introduce a more positive tone. Across all models, the increase in positive sentiment consistently exceeds the corresponding decrease in negative sentiment, suggesting that refinement models do more than simply filter out negativity.

To illustrate this transformation, consider the following excerpts from a pricing negotiation generated by Llama-8B and refined by GPT-4. In the base generation, the exchange escalates in tone: the counterparty responds to an initial offer with "Are you asking me if I want to do the job or not? I will lift the diesel at US$75 per metric ton" and later reacts to a lower price with "Are you insane? I will lose money if I lift at US$60 per metric ton." After refinement, these become "Are you asking me to confirm the job? I would propose lifting the diesel at US$75 per metric ton instead" and "Lifting at US$60 per metric ton would not be financially viable for us. We might need to explore alterna-

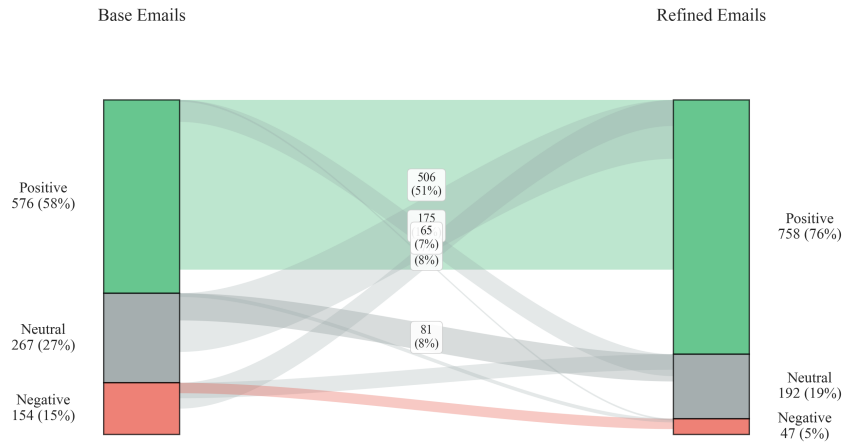

Figure 2: Sentiment transitions during GPT-4's refinement process in the BARE pipeline. The left column shows the sentiment distribution of the base-generated emails (Llama-8B), and the right column shows the distribution after refinement. Flow widths are proportional to the number of emails undergoing each sentiment transition and indicate that 69% of initially negative emails shift to either neutral or positive sentiment.

tive arrangements if the price cannot be adjusted higher." The confrontational tone, typical of real freight negotiations, is systematically replaced by diplomatic hedging and more collaborative framing.

Negative sentiment is underrepresented in our synthetic emails relative to the Enron reference corpus and likely also relative to many real negotiation settings, where frustrations over pricing, disagreements about contractual terms, and assertive language are common. This systematic positivity likely stems from the safety training and instruction-tuning procedures of modern LLMs, which emphasize helpful, harmless, and honest outputs (Askell et al., 2021; Bai et al., 2022). These tendencies have also been linked to both instruction-tuning objectives and recursive model-training effects that can amplify stylistic bias (Li et al., 2025; Shumailov et al., 2024). While these objectives are appropriate for general applications, they may limit the models' ability to generate typical business communication that includes a more direct or confrontational tone. As a result, the dataset tends to reflect a more collaborative negotiation style, which could lead models trained on it to underrepresent urgency or disagreement in real exchanges.

5. Extrinsic Evaluation: Information Extraction Baselines

To demonstrate the dataset's dual role as both a training resource and an evaluation benchmark, we establish baselines for entity extraction using two complementary approaches targeting four entity types: vessel names, ports or locations, commodities, and Incoterms. As described in Section 3, these annotations were produced by the generating models during synthesis (or by the refining model in the BARE setting); they therefore serve as machine-generated reference labels for the experiments below.

First, we apply GLiNER (Zaratiana et al., 2023), a span-based architecture that jointly encodes input text and entity type labels to identify entity boundaries. We test several encoder pretraining strategies, including entity-contrastive learning, SimCSE (Gao et al., 2021), and masked language modeling, combined with different domain adaptation approaches. The best configuration uses memory-bank contrastive fine-tuning, which maintains separate memory banks for vessels and locations to improve disambiguation between overlapping entity types (e.g., distinguishing the vessel *MV Singapore* from the port *Singapore*). We also fine-tune Qwen-2.5-0.5B-Instruct (Yang et al., 2024) using parameter-efficient fine-tuning (LoRA), formulating entity extraction as a JSON generation task. This compact generative approach performs competitively and supports flexible output formats beyond the predefined entity types.

Performance Results Overview Figure 3 shows macro F1 scores across data generation methods and fine-tuning configurations. The results reveal consistent patterns. GLiNER consistently achieves higher and more stable performance across all

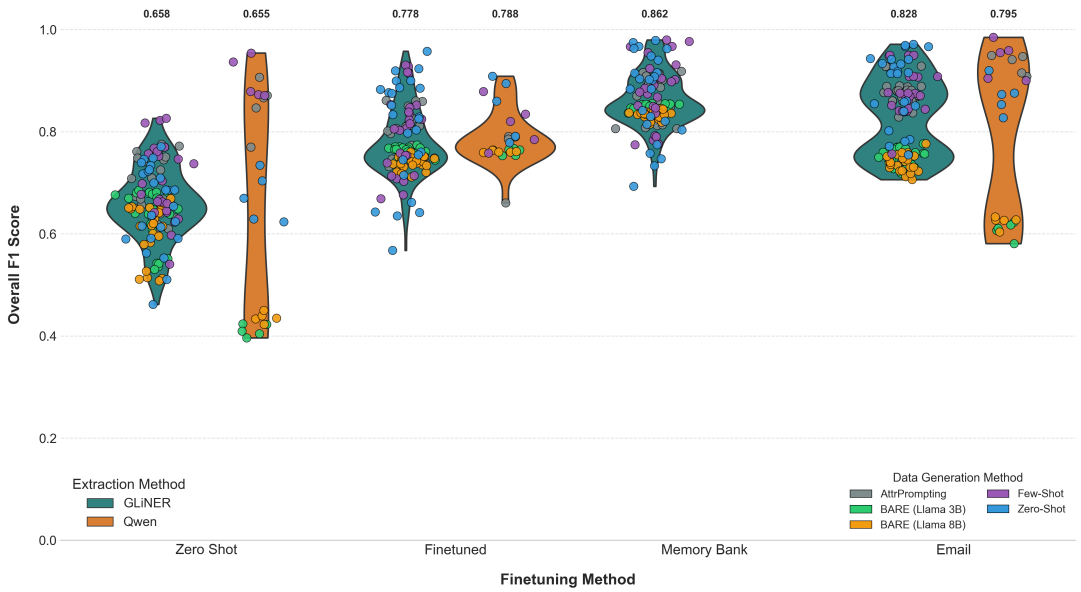

Figure 3: Overall F1 score comparison between GLiNER and Qwen across different experimental conditions. The x-axis shows the fine-tuning settings: **Zero Shot** (no fine-tuning), **Fine-tuned** (general maritime corpus), **Memory Bank** (GLiNER-specific contrastive fine-tuning), and **Email** (fine-tuning on the target email distribution). Each dot represents performance on a dataset generated with a specific data generation configuration, with colors indicating the generation method. GLiNER consistently shows higher and more stable performance, with the Memory Bank approach achieving the highest overall score (0.862). Qwen's performance is more variable and depends more strongly on the match between fine-tuning data and evaluation data.

settings, with the memory-bank contrastive variant reaching the best overall F1 score (0.862). Qwen shows greater variability and depends more strongly on the match between fine-tuning data and evaluation data. Extraction difficulty appears to increase with generation diversity: BARE-generated emails, which show the greatest linguistic variety, are also the most challenging for entity extraction, especially for the generative Qwen model. This pattern suggests a trade-off in which more diverse training data increases extraction difficulty while potentially improving robustness to linguistically variable and domain-plausible text.

Entity-level results reveal substantial variation in extraction difficulty across entity types. Incoterms are the hardest to identify, with F1 scores 20–30 points lower than those for other entity types. These short three-letter trade terms (e.g., CIF, FOB) frequently resemble other maritime abbreviations, and only models exposed to comprehensive maritime terminology during fine-tuning can reliably distinguish them. The two architectures display complementary strengths: GLiNER offers greater consistency and efficiency (1,000 emails / minute), whereas Qwen provides more flexible extraction formats but slower inference (60 emails / minute) and higher variance. The memory-bank contrastive variant again performs best (0.862 F1), suggesting that explicit contrastive learning helps address entity disambiguation in the maritime domain. Fine-tuned models exceed 0.85 F1 on more structured generation settings, while the 20-point absolute performance gap between zero-shot and fine-tuned runs highlights the value of domain-specific training data.

6. Limitations and Future Work

6.1. Limitations

This study has several limitations that should be considered when interpreting the results. First, hardware limitations prevented the use of larger base models for BARE generation. While Llama-8B produced diverse outputs, larger models may generate more domain-plausible maritime correspondence and could affect downstream extraction difficulty. Second, the Enron corpus, our only human-written baseline, has known limitations due to its age, its general business focus, and its widespread use in model pretraining.

Additionally, resource constraints prevented exhaustive hyperparameter optimization for all model configurations. The relative performance differences between extraction methods might shift under different settings, and some models may not

have reached their best performance. Finally, the entity labels were produced by the generating models themselves (or, in the BARE setting, by the refining model) rather than by human annotators. Extraction performance therefore reflects agreement with machine-generated annotations, which may contain inconsistencies—particularly for BARE outputs, where the refining model must infer entities from loosely structured base text. This design choice, driven by the scale of the corpus, means that the reported F1 scores may overestimate performance relative to what manual evaluation would yield.

6.2. Future Work

Future research can address several of these limitations. First, scaling experiments with larger base models (70B parameters or more) may improve the domain plausibility and terminological accuracy of generated emails. Such models could better capture the specialized terminology and negotiation style typical of maritime communication, potentially narrowing the gap between synthetic and human-written correspondence observed in our diversity analyses. In addition, recent methods such as hindsight merging (Veselovsky et al., 2025) show promise for improving the balance between diversity and coherence in synthetic text generation. Complementary approaches such as generalized instruction tuning (Li et al., 2024, 2023) could further expand controllable diversity while retaining factual consistency.

Second, domain-specific pretraining of base models on maritime corpora is a promising approach for improving terminological accuracy and domain plausibility. Curated pretraining data from maritime news, shipping bulletins, and industry publications could help models internalize domain conventions more effectively. Addressing the systematic positivity bias observed in our sentiment analysis will also require targeted strategies. Future work could explore specialized prompting, reward modeling, or post-hoc filtering to produce more plausible sentiment distributions and more direct negotiation language without compromising safety objectives.

Finally, expanding the evaluation beyond entity extraction would provide a broader assessment of the dataset's utility. Potential downstream tasks include email-thread classification, negotiation-outcome prediction, sentiment-aware response generation, and automated contract-term extraction. Additionally, although our evaluation focused on locally hosted models for reproducibility and cost efficiency, benchmarking against state-of-the-art commercial APIs could help establish an approximate upper bound for maritime information extraction performance. Such evaluations must, however, consider privacy carefully when processing potentially sensitive business communication.

7. Conclusion

This paper introduced MaritimEmails, a synthetic corpus of 19,817 email threads designed to address the lack of publicly available datasets for maritime business correspondence. Using multiple generation strategies across five contemporary language models, the dataset provides a diverse resource for developing domain-specific NLP applications in the maritime sector. Our evaluation shows that BARE approaches yield substantially higher lexical diversity than standard prompting, with the best configuration (Llama-8B with Claude refinement) approaching the Enron reference corpus in phrasal diversity while producing coherent outputs.

At the same time, the analysis reveals a systematic positivity bias in current language models: synthetic emails rarely express the level of negative sentiment observed in the Enron reference corpus and likely also underrepresent disagreement in many real negotiation settings. This finding reflects a broader tension between model safety objectives and the goal of generating domain-plausible business communication. Nevertheless, baseline experiments show that models trained on MaritimEmails achieve F1 scores above 0.85 for entity extraction, indicating the dataset's usefulness for text mining tasks. The trade-off between generation diversity and extraction difficulty suggests that BARE-generated emails, while harder to process, may better prepare models for linguistically variable and domain-plausible text.

The dataset, including structured entity annotations for vessels, ports, commodities, and Incoterms, is publicly available at `https://github.com/ZurichNLP/MaritimEmails` under CC BY-NC 4.0.

Ethical and Practical Considerations All data in MaritimEmails are fully synthetic and contain no personal or real business information. Nonetheless, the generated texts may reproduce stylistic or sentiment biases present in the underlying language models. Each record includes a `"synthetic": true` flag to prevent misattribution, and the CC BY-NC 4.0 license restricts commercial reuse. Researchers are encouraged to disclose the synthetic provenance of the data when using it in downstream studies.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. ArXiv:2303.08774 [cs.CL].

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. ArXiv:2112.00861 [cs.CL].

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. ArXiv:2204.05862 [cs.CL].

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. ArXiv:2410.15226 [cs.CL].

Parag Pravin Dakle and Dan Moldovan. 2020. CEREC: A corpus for entity resolution in email conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 339–349, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. ArXiv:2407.21783 [cs.CL].

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. ArXiv:2412.10271 [cs.CL].

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. ArXiv:2311.09807 [cs.CL].

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. 2024. Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. ArXiv:2412.00586 [cs.CR].

Ibrahim. 2023. Global Cargo Ships Dataset. Kaggle Dataset. [CC0: Public Domain], accessed 10 March 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. ArXiv:2401.04088 [cs.LG].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. ArXiv:2001.08361 [cs.LG].

Rabimba Karanjai. 2022. Targeted phishing campaigns using large scale language models. ArXiv:2301.00665 [cs.CR].

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*. Mountain View, California, USA, July 30–31, 2004.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation. ArXiv:2407.11016 [cs.CL].

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. ArXiv:2402.13064 [cs.CL].

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. ArXiv:1510.03055 [cs.CL].

Weijiang Li, Yinmeng Lai, Sandeep Soni, and Koustuv Saha. 2025. Emails by LLMs: A comparison of language in AI-Generated and human-written emails. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 391–403, New York, NY, USA. Association for Computing Machinery.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. ArXiv:2412.19437 [cs.CL].

Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when i lose my pet? understanding perceptions of AI-mediated email writing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024a. On llms-driven synthetic data generation, curation, and evaluation: A survey. ArXiv:2406.15126 [cs.CL].

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024b. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2023. Machine learning for synthetic data generation: A review. ArXiv:2302.04062 [cs.LG].

Mihai Nadăş, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

Sanjeet Singh Naik. 2023. Shipping Ports Around The World. Kaggle Dataset. [CC0: Public Domain], accessed 10 March 2025.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? ArXiv:2309.05196 [cs.CL].

G. Papageorgiou, P. Economou, and S. Bersimis. 2024. A method for optimizing text preprocessing and text classification using multiple cycles of learning with an application on shipbrokers emails. *Journal of Applied Statistics*, 51(13):2592–2626.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Philippe Remy. 2021. Name dataset. GitHub repository.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. ArXiv:2403.00553 [cs.CL].

Sung-Ho Shin, Oh Kyoung Kwon, Xiao Ruan, Prem Chhetri, Paul Tae-Woo Lee, and Shahrooz Shahparvari. 2018. Analyzing sustainability literature in maritime studies with text mining. *Sustainability*, 10(10):3522.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Saurabh Srivastava, Gaurav Singh, Shou Matsumoto, Ali Raz, Paulo Costa, Joshua Poore, and Ziyu Yao. 2023. Mailex: Email event and argument extraction. ArXiv:2305.13469 [cs.CL].

Statista Research Department. 2024. Daily number of e-mails worldwide from 2017 to 2025. Accessed: 2025-10-15.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. ArXiv:2312.11805 [cs.CL].

United Nations Conference on Trade and Development. 2024. Review of maritime transport 2024. Technical report, UNCTAD.

Veniamin Veselovsky, Benedikt Stroebl, Gianluca M. Bencomo, Dilip Arumugam, Lisa Schut, Arvind Narayanan, and Thomas L. Griffiths. 2025. Hindsight merging: Diverse data generation with language models. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 4349–4369.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. ArXiv:2407.10671 [cs.CL].

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, Datasets and Benchmarks Track*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist model for named entity recognition using bidirectional transformer. ArXiv:2311.08526 [cs.CL].

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909.

Yusheng Zhou, Xueqin Wang, and Kum Fai Yuen. 2021. Sustainability disclosure for container shipping: A text-mining approach. *Transport Policy*, 110:465–477.

Alan Zhu, Parth Asawa, Jared Quincy Davis, Lingjiao Chen, Ion Stoica, Joseph E. Gonzalez, and Matei Zaharia. 2025. BARE: Combining base and instruction-tuned language models for better synthetic data generation. ArXiv:2502.01697 [cs.CL].