Zero-Shot LLM Generation of Energy Notifications for African Languages: A Benchmark Study

Hatem Haddad hatem.haddad@wattnow.io
Feres Jerbi feres.jerbi@wattnow.io
Issam Smaali issam@wattnow.io
Wattnow, Tunisia

Abstract

Large Language Models (LLMs) have demonstrated significant advancements in natural language applications but often exhibit performance disparities for low-resource languages, particularly African languages underrepresented in training corpora. This paper addresses this gap by evaluating the zero-shot text generation capabilities of LLMs within the energy domain for six widely spoken African languages. We introduce a novel multilingual benchmark dataset of energy management notifications and use it to assess four recent open-source LLMs (1B-7B parameters). Employing a zero-shot learning approach with multiple prompts and established NLP metrics (Statistics-based, Model-based, Perplexity) without fine-tuning, our findings reveal varying model strengths across languages and metrics. For instance, while some models excel in content overlap (ROUGE) for languages like English and French, others show better fluency (Perplexity) or semantic similarity (BERTScore), with performance shifting notably for African languages.

Keywords: Large Language Models (LLMs), Zero-Shot Learning, Text Generation, African Languages, Energy Management

1. Introduction

The recent surge in popularity of Large Language Models (LLMs), particularly since ChatGPT's release, has generated significant excitement for their potential applications across numerous fields. Indeed, LLMs have achieved remarkable progress in Natural Language Processing and beyond, demonstrating sophisticated capabilities in understanding and generating text Wan et al. (2023). Text generation, a crucial computational task focused on creating human-like text Becker et al. (2024), is central to these advancements, enhancing accessibility, writing productivity, and cross-lingual communication Ramos et al. (2023), though it encompasses diverse methods beyond just autoregressive models Katz (1980). LLMs are increasingly integrated into diverse real-world applications, facilitating complex tasks from automatic code generation and powering interactive chatbots to enabling advanced data analysis. For instance, recent work has demonstrated LLMs effectively automating clinical note generation from doctor-patient dialogues, significantly reducing documentation burdens and improving note quality Li et al. (2024). Similarly, other studies have deployed LLMs to generate dynamic, domain-specific narratives from structured data in sectors such as sports and music, showcasing high fidelity and large-scale fan engagement Baughman et al. (2024).

Despite these successes, LLM performance often excels primarily in high-resource languages where abundant training data is available Minaee et al. (2024). Significant disparities persist for low-resource languages, especially for the vast diversity of native African languages constituting approximately one-third of the world's languages which remain severely underrepresented in common training datasets Buzaaba et al. (2025); Lawal et al. (2024); Alhanai et al. (2025); Aryabumi et al. (2024). This underrepresentation leads to notably poorer performance in text generation tasks, as exemplified by studies on transformer-based models for languages like Nigerian Pidgin, which reported challenges in achieving high fluency and accuracy Garba et al. (2024).

This paper addresses the critical performance gap of Large Language Models (LLMs) for African languages within specialized domains, particularly energy text generation. To the best of our knowledge, this work represents the first comprehensive benchmarking of LLMs in this specific context. We introduce a novel multilingual dataset comprising energy management notifications meticulously curated for six widely spoken African languages, which forms the basis of our investigation. On this benchmark, we evaluate four recent open-source LLMs (ranging from 1B to 7B parameters) using a zero-shot learning approach; this involves multiple prompts for eight distinct notification types per language and assessment with established NLP metrics (Statistics-based, Model-based, and Perplexity) without any model fine-tuning. The resulting findings are intended to guide the development of more equitable and effective language technologies.

2. Energy Management Notification

Energy Management Notifications (EMNs) provide customers with energy usage insights and optimization strategies to enhance awareness, efficiency, and demand-side participation, ultimately aiming for cost savings and grid stability Schwartz et al. (2015). However, manual EMN delivery is often impaired by critical inefficiencies, error-proneness, high operational costs, and an inability to offer timely, personalized information, thereby hindering these objectives. Consequently, automated systems leveraging text generation are proposed to overcome such limitations, offering superior operational efficiency, improved accuracy, real-time capabilities, and scalable personalization to effectively meet EMN goals, reduce costs, and boost customer engagement Buchanan et al. (2015).

Essential data for the automated energy notification system encompassing customer profiles, consumption metrics, and notification parameters is securely hosted on a cloud service. For evaluation, eight notifications per language were utilized as references (Ground Truth). These notifications were manually crafted by a domain expert, specifically an energy engineer with expertise in product-customer interactions, covering six distinct topics such as Subscribed Power.

Table 1: Energy Data Used to Generate Subscribed Power Notification

Type	Measure	Unit	Value	Threshold	Percentage	Detection Time
Electricity	Subscribed Power	kW	606.25	600	75	2025-03-11 06:16:51

Table 1 outlines the features used for generating the Subscribed Power notification: Type (system category), Measure (monitored parameter, e.g., Subscribed Power), Unit (e.g., kW), Value (observed measurement), Threshold (reference limit), Percentage (client-defined margin), Threshold Type (nature of breach), and Detection Time (timestamp). In this example, an alert was triggered when Subscribed Power reaches 75% of the threshold of 600 kW, as defined by the client. This data is used as the LLM input on a structured JSON file to craft personalized customer messages as shown in Table 2.

3. Evaluated Models & Languages

For the evaluation, we selected four models, based on two primary criteria. First, each model is openly available and licensed for commercial use, aligning with our focus on supporting deployment within a startup ecosystem. Second, we prioritized small-sized models to ensure faster inference and lower computational costs, enabling practical deployment on edge devices and in resource-constrained environments. For these reasons, four LLM models based on Transformer decoder-only architecture are evaluated. Qwen2.5:3B (referred to as Qwen from now on), by Alibaba, last updated on May 28, 2025, has 3 billion parameters and was trained on 18 trillion tokens, with a disk size of 1.9 GB. Mistral:7B (referred to as Mistral from now on), developed by Mistral AI and last updated on May 22, 2024, contains 7 billion parameters, though its training dataset remains undisclosed, and occupies 4.4 GB. DeepSeek-R1:1.5b (referred to as DeepSeek from now on), released by DeepSeek on May 2, 2025, has 1.5 billion parameters and a 1.1 GB footprint, with no details about its training dataset. Meta's Llama3.2:1B (referred to as Llama from now on), last updated on September 25, 2024, includes 1 billion parameters and takes up 1.3 GB, also with unspecified training data.

This study focuses on a diverse set of languages. English, from the Anglophone region and the Indo-European family, is spoken by 237 million African people. French, also an Indo-European language from the Francophone region, has 320 million African speakers. Arabic, Semitic language, is spoken by 275 million individuals in North Africa. Representing the Bantu languages, Swahili, from the East Bantu branch, has 16 million speakers, while IsiZulu, a South Bantu language, is spoken by 13.6 million people. Finally, Hausa, a West Chadic language, is spoken by 54 million people.

Table 2: Manually Generated Ground Truth 'Subscribed Power' Notifications in Six Languages (Based on Data from Table 1)

French: Cher Client, Nous avons détecté un dépassement de la puissance souscrite pour le site **X**. La valeur mesurée a atteint 606,25 kW, dépassant ainsi la puissance prévue dans votre contrat. Veuillez noter que tout dépassement prolongé au-delà de 10 minutes peut entraîner des pénalités sur votre facture énergétique. **Actions immédiates :** Vérifier les équipements énergivores : Identifiez les équipements responsables de cette consommation excessive. Analyser les charges actives : Examinez les équipements actuellement en fonctionnement pour déterminer la source de cette augmentation inattendue. Inspecter le système électrique : Vérifiez le système électrique afin d'exclure toute anomalie. **Actions recommandées :** Analyser les pratiques opérationnelles : Évaluez si une modification du contrat de puissance ou des ajustements opérationnels sont nécessaires pour éviter de futurs dépassements. Pour obtenir plus d'informations ou une assistance complémentaire, veuillez consulter notre tableau de bord ou contacter directement notre équipe de support technique.	***English:***Dear Customer, We have detected an exceedance of the subscribed power for the **X** site. The recorded value reached 606.25 kW, surpassing your contractual subscribed power. Please note that any prolonged exceedance beyond 10 min may result in penalties on your energy bill. **Immediate Actions:** Check High-Consumption Equipment: Identify the equipment responsible for this excessive consumption. Analyze Active Loads: Examine the equipment currently operational to pinpoint the source of this unexpected increase. Inspect Electrical System: Verify the electrical system to rule out any anomalies. **Recommended Actions:** Review Operational Practices: Assess if adjusting your power subscription or operational practices is necessary to prevent future exceedance. For additional information or further assistance, please visit our dashboard or directly contact our technical support team.
Arabic: عزيزي العميل، لقد اكتشفنا وجود فائض في الطاقة المشتركة لموقع خ. بلغت القيمة المقاسة ٢٥٠٦٠٦ كيلو وات، وهو ما يتجاوز الطاقة المنصوص عليها في عقدك. يرجى ملاحظة أن أي تجاوز لمدة تزيد عن ١٠ دقائق قد يؤدي إلى فرض عقوبات على فاتورة الطاقة الخاصة بك. الإجراءات الفورية: فحص المعدات المستهلكة للطاقة: حدد المعدات المسؤولة عن هذا الاستهلاك المفرط. تحليل الأحمال النشطة: فحص المعدات العاملة حاليًا لتحديد مصدر هذه الزيادة غير المتوقعة. فحص النظام الكهربائي: فحص النظام الكهربائي لاستبعاد أي خلل. الإجراءات الموصى بها: تحليل الممارسات التشغيلية: تقييم ما إذا كان من الضروري إجراء تعديل على عقد الطاقة أو تعديلات تشغيلية لتجنب التجاوزات في المستقبل. لمزيد من المعلومات أو المساعدة الإضافية، يرجى زيارة لوحة التحكم الخاصة بنا أو الاتصال بفريق الدعم الفني لدينا مباشرة.	***isiZulu:*** Mthengi Ohloniphekile, Siqaphele ukuthi amandla okubhaliselwe esizeni **X** kudlule. Inani elirekhodiwe lafika ku-606.25 kW, lidlula amandla okubhaliselwe okufanele. Sicela uqaphele ukuthi noma yikuphi ukweqa okuphinde kube isikhathi eside kune-10 imizuzu kungaholela kumaholo okwephula umthetho kumabhili wakho wamandla. **Izinyathelo Ezilungile:** Bheka Izisetshenziswa Ezisebenzisa Amandla Amaningi: Thola izisetshenziswa ezibangela lokhu kudla amandla ngokweqile. Hlaziya Umthwalo Osebenzayo: Hlola izisetshenziswa ezisebenzayo njengamanje ukuze uthole imbangela yalokhu kukhuphuka okungahleleki. Bheka Uhlelo Lwamandla: Qinisekisa uhlelo lwezokuhlinzeka ngamandla ukuze uqinisekise ukuthi akukho okungahambanga kahle. **Izinyathelo Ezilungile Ezilandelayo:** Bheka Izindlela Zokusebenza: Hlola ukuthi kumele yini ushintshe ukubhalisa kwakho kwamandla noma izindlela zakho zokusebenza ukuze ugweme ukweqa okwengeziwe. Uma udinga eminye imininingwane noma usizo oluthe xaxa, sicela uvakashele i-dashboard yakho noma uxhumane ngqo nethimba lethu lokweseka kwezobuchwepheshe.
Swahili: Mteja Mpendwa, Tumeona kuwa matumizi ya umeme katika kituo cha **X** yamezidi kiwango cha uwezo wa umeme ulicholipiwa. Thamani iliyorekodiwa imefikia 606.25 kW, ambayo imezidi kiwango chako cha mkataba. Tafadhali fahamu kuwa ukiukaji huu ukidumu kwa zaidi ya dakika 10 unaweza kusababisha adhabu kwenye bili yako ya umeme. **Hatua za Haraka:** Kagua Vifaa Vinavyotumia Umeme kwa Kiasi Kikubwa: Tambua kifaa kinachosababisha matumizi haya makubwa. Chambua Mizigo Inayotumika: Kagua vifaa vinavyofanya kazi kwa sasa ili kubaini chanzo cha ongezeko hili lisilotarajiwa. Kagua Mfumo wa Umeme: Hakikisha mfumo wa umeme hauna hitilafu au tatizo lolote. **Hatua Zinazopendekezwa:** Pitisha upya Taratibu za Uendeshaji: Angalia kama kuna haja ya kurekebisha usajili wa uwezo wa umeme au kubadili mbinu zako za uendeshaji ili kuepuka uongezekaji kama huu siku zijazo. Kwa maelezo zaidi au msaada wa ziada, tafadhali tembelea dashibodi yetu au wasiliana moja kwa moja na timu yetu ya msaada wa kiufundi.	***Hausa:*** Abokin Ciniki Mai Daraja, Mun gano cewa an wuce karfin wutar lantarki da aka yi rajista a shafin **X**. An samu darajar 606.25 kW, wanda ya wuce karfin wutar lantarki da aka yi rajista a cikin kwangilar ku. Da fatan za a lura cewa kowanne lokaci mai tsawo na wucewa fiye da mintuna 10 zai iya haifar da tara a kan lissafin ku na makamashi. **Matakan Gaggawa:** Duba Kayan Aikin da Ke Cinye Wuta Mai Yawa: Tantance kayan aikin da ke haifar da wannan yawan amfani. Bincika Nau'ukan Aiki: Duba kayan aikin da ke aiki a yanzu don gano tushen wannan karin amfani da ba a zata ba. Duba Tsarin Lantarki: Tabbatar da tsarin lantarki don tabbatar da cewa babu wani abu da ba a saba da shi ba. **Matakan da Aka Ba da Shawara:** Duba Ayyukan Gudanarwa: Tabbatar ko canza rajistar wutar lantarki ko yin gyara a cikin ayyukan ku zai taimaka wajen kauce wa wucewar karfi a nan gaba. Don arin bayani ko taimako, da fatan za a ziyarci dashboard inmu ko tuntui ungiyar tallafin fasaha.

Table 3: Average Notification Word Count by Language and Model

Language	Reference	Qwen	Mistral	DeepSeek	Llama
English	136.63	109.82	136.50	130.91	109.82
French	156.50	189.22	179.89	151.63	146.11
Arabic	135.00	185.71	141.25	216.00	136.88
Swahili	155.88	185.00	261.88	280.13	242.25
Hausa	171.75	139.88	186.13	224.36	109.13
isiZulu	113.75	135.00	153.25	162.13	109.25

4. Evaluation Metrics

We assessed lexical similarity between generated and reference notification using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics Lin (2004). Specifically, ROUGE-1 evaluated unigram overlap, ROUGE-2 assessed bigram overlap, and ROUGE-L identified the longest common subsequence (LCS), all quantifying content fidelity at the token level by comparing candidate and reference notification. We employed BERTScore Zhang et al. (2019) as a model-based metric for a nuanced understanding of semantic similarity, going beyond lexical overlap. Known for its robust correlation with human perception, BERTScore assesses content alignment by computing pairwise cosine similarity between contextual BERT embeddings of generated and reference notification tokens

Perplexity Cooper and Scholak (2024) was employed to assess the linguistic coherence and fluency of the generated texts. It quantifies a language model's uncertainty in predicting a sequence of tokens, with lower perplexity scores indicating greater model confidence and more natural, fluent output. Essentially, lower perplexity reflects how well the language model "understands" the linguistic structure of the evaluated text.

5. Performances Results and Discussion

This section presents the best-performing results achieved after testing various prompts for each language. The mean evaluation metric scores for all generated notifications per language are detailed in Tables 4, 5, 6, 7, 8 and 9, where the highest scores are highlighted in bold. Concurrently, Table 3 details the average word count for both reference notifications and their generated counterparts using the default temperature parameter of 0.8. It reveals critical disparities in LLM-generated notification lengths: models like DeepSeek and Qwen were excessively verbose in Arabic, whereas for Hausa and isiZulu varied, with some models producing longer and others shorter texts than references. These inconsistencies in length across languages highlight challenges in controlling output verbosity, potentially impacting information completeness or conciseness in generated alerts.

Table 4: English Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	**0.44**	0.43	0.31	0.34
ROUGE-2	**0.16**	0.13	0.08	0.09
ROUGE-L	**0.27**	0.23	0.20	0.21
BERTScore	**0.69**	0.67	0.63	0.62
Perplexity	40.36	**33.17**	56.87	46.95

Table 5: French Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	**0.40**	0.38	0.35	0.35
ROUGE-2	**0.11**	0.11	0.07	0.077
ROUGE-L	**0.23**	0.21	0.11	0.22
BERTScore	**0.65**	0.62	0.59	0.60
Perplexity	63.49	**43.24**	68.74	72.28

Our findings, based on ROUGE scores, BERTScore, and Perplexity, indicate that no single model universally outperforms others across all languages and metrics, highlighting the complex interplay between model architecture, training data, and linguistic characteristics. Across English and French (Tables 4 and 5), Qwen consistently achieved the highest ROUGE-1, ROUGE-2, and BERTScore. This suggests a strong capability in capturing n-grammatic overlap and semantic similarity with the reference texts for these relatively high-resource languages. For generating notifications, high ROUGE scores are desirable as they often correlate with the inclusion of key factual details from the input data (the abnormal energy usage patterns). The high BERTScore further reinforces its ability to produce semantically coherent and relevant content, which is crucial for ensuring clients understand the alert. However, in both English and French, Mistral, despite slightly lower ROUGE and BERTScore values than Qwen, exhibited markedly lower Perplexity. This indicates that Mistral generates more fluent and predictable text in these languages. The slightly larger size of Mistral:7b parameters vs. Qwen:3b might contribute to its enhanced fluency, potentially stemming from a more extensive general language understanding.

The performance landscape shifts for Arabic and the African languages, underscoring the challenges LLMs face with languages potentially underrepresented in their training corpora. For Arabic, Qwen again led in ROUGE scores, though the absolute values are notably lower than in English or French, suggesting greater difficulty in matching reference

Table 6: Arabic Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	**0.14**	0.12	0.04	0.11
ROUGE-2	**0.07**	0.06	0.02	0.05
ROUGE-L	**0.30**	0.22	0.04	0.12
BERTScore	0.72	**0.74**	0.69	0.70
Perplexity	9.16	11.00	10.36	**9.01**

Table 7: isiZulu Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	0.13	**0.13**	0.05	0.07
ROUGE-2	0.03	**0.04**	0.02	0.02
ROUGE-L	0.09	**0.09**	0.05	0.05
BERTScore	0.67	**0.67**	0.50	0.66
Perplexity	73.10	80.60	**18.20**	28.35

phrasing. Conversely, Mistral achieved the best BERTScore and Llama achieved the lowest Perplexity slightly ahead of Qwen model, indicating better semantic relevance and fluency.

For isiZulu (Table 7), Mistral secured the best ROUGE scores and achieved the highest BERTScore. Surprisingly, the much smaller DeepSeek recorded the lowest Perplexity by a substantial margin, suggesting it generated the most predictable, and potentially fluent, text despite weaker content overlap. This could imply that DeepSeek might be well-tuned for the structural aspects of isiZulu but struggles with content fidelity for this specific task. Mistral's high Perplexity here is also noteworthy and warrants further investigation. For Swahili (Table 8), Qwen continued its ROUGE and BERTScore dominance. However, the DeepSeek model achieved the lowest Perplexity. This is a compelling result, it suggests that for DeepSeek, might offer the best balance of semantic accuracy and fluency for the given task, outperforming larger models in these aspects. For Hausa (Table 9), Mistral demonstrated clear superiority in all ROUGE metrics. Llama showed a strong BERTScore, while DeepSeek, similar to its performance in isiZulu and Hausa, achieves a low Perplexity. This pattern with DeepSeek in certain African languages (very low perplexity but weaker ROUGE/BERTScore) might indicate a tendency towards generating structurally simple or repetitive, albeit fluent, text that doesn't fully capture the source data's nuances.

The evaluation of LLMs for generating energy notifications highlights specific model strengths and trade-offs crucial for this task. Qwen consistently demonstrates strong content overlap (ROUGE scores) across multiple languages, making it a good candidate for

Table 8: Swahili Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	**0.28**	0.26	0.09	0.26
ROUGE-2	**0.06**	0.05	0.02	0.05
ROUGE-L	**0.17**	0.16	0.07	0.15
BERTScore	**0.66**	0.64	0.48	0.65
Perplexity	87.63	63.96	**31.63**	46.03

Table 9: Hausa Performances

Metric	**Qwen**	**Mistral**	**DeepSeek**	**Llama**
ROUGE-1	0.27	**0.35**	0.03	0.16
ROUGE-2	0.08	**0.11**	0.01	0.03
ROUGE-L	0.15	**0.17**	0.05	0.13
BERTScore	0.59	0.61	0.47	**0.64**
Perplexity	73.27	42.34	**6.38**	73.03

factual accuracy, though its fluency might require improvement via prompting or fine-tuning. Conversely, Mistral often excels in fluency (low Perplexity), which is vital for client understanding. Ideally, a model should combine high ROUGE/BERTScore (for factual and semantic accuracy) with low Perplexity (for natural delivery).

6. Conclusion

This study benchmarked four open-source LLMs for zero-shot energy notification generation across six African languages and revealed significant, language- and metric-dependent performance variations, with no single model universally excelling. While models like Qwen demonstrated strong content overlap in English and French, DeepSeek proved competitive in terms of perplexity for isiZulu, Swahili, and Hausa. This evaluation underscores the necessity of language-specific model selection and highlights the viability of efficient smaller models for specialized, low-resource domains. Our study intentionally focused on zero-shot performance as an initial exploration into data-to-text generation within the energy sector. Future research will expand upon this by evaluating and comparing zero-shot outcomes with fine-tuned models and retrieval-augmented generation (RAG) architectures. These planned comparisons will help clarify the trade-offs between generalization and task-specific optimization. Ultimately, the dataset and findings presented here aim to foster more equitable language technologies tailored specifically for African languages and contexts.

References

T. Alhanai, A. Kasumovic, M. M. Ghassemi, A. Zitzelberger, J. M. Lundin, and G. Chabot-Couture. Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812, 2025.

V. Aryabumi, J. Dang, D. Talupuru, S. Dash, D. Cairuz, H. Lin, B. Venkitesh, M. Smith, K. Marchisio, S. Ruder, A. F. Locatelli, J. Kreutzer, N. Frosst, P. Blunsom, M. Fadaee, A. Ustun, and S. Hooker. Aya 23: Open weight releases to further multilingual progress. *arXiv*, 2405.15032, 2024.

A. Baughman, E. Morales, R. Agarwal, G. Akay, R. Feris, T. Johnson, S. Hammer, and L. Karlinsky. Large scale generative ai text applied to sports and music. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4784–4792, 2024.

J. Becker, J. P. Wahle, B. Gipp, and T. Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv*, 2405.15604, 2024.

K. Buchanan, R. Russo, and B. Anderson. The question of energy reduction: The problem(s) with feedback. *Energy Policy*, 77:89–96, 2015.

H. Buzaaba, A. Wettig, D. I. Adelani, and C. Fellbaum. Lugha-llama: Adapting large language models for african languages. *arXiv*, 2025. arXiv e-prints (ID: arXiv–2504).

N. Cooper and T. Scholak. Perplexed: Understanding when large language models are confused. *arXiv*, 2404.06634, 2024.

K. Garba, T. Kolajo, and J. B. Agbogun. A transformer-based approach to nigerian pidgin text generation. *International Journal of Speech Technology*, pages 1–11, 2024.

B. Katz. A three-step procedure for language generation. 1980.

O. I. Lawal, O. Adekanmbi, and A. Soronnadi. Contextual evaluation of llm's performance on primary education science learning contents in the yoruba language. In *5th Workshop on African Natural Language Processing*, 2024.

Y. Li, S. Wu, C. Smith, T. Lo, and B. Liu. Improving clinical note generation from complex doctor-patient conversation. *arXiv*, 2408.14568, 2024.

C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

S. Minaee, T. Mikolov, N. Nikzad, M. Asgari Chenaghlu, R. Socher, X. Amatriain, and J. Gao. Large language models: A survey. *arXiv*, 2402.06196, 2024.

L. Ramos, R. Márquez, and F. Rivas-Echeverría. Ai's next frontier: The rise of chatgpt and its implications on society, industry, and scientific research la próxima frontera de la ia: El surgimiento de chatgpt y sus implicaciones en la sociedad, la industria y la investigación científica. *Revista Ciencia e Ingeniería*, 44(2), 2023.

D. K. Schwartz, B. Fischhoff, G. Loewenstein, and E. U. Weber. What information do people want about energy use? *Energy Policy*, 80:169–179, 2015.

Z. Wan, X. Wang, C. Liu, S. Alam, Y. Zheng, J. Liu, Z. Qu, S. Yan, Y. Zhu, Q. Zhang, et al. Efficient large language models: A survey. *arXiv*, 2312.03863, 2023.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv*, 1904.09675, 2019.