Arabic ChartSumm: An English-to-Arabic Benchmark for Metadata-to-Text Summarization

Passant Elchafei[1,*] **Amany Fashwan**[2,*]
[1]Ulm University, Germany
[2]Alexandria University, Egypt
passant.elchafei@uni-ulm.de, amany.fashwan@alexu.edu.eg

Abstract

Generating summaries from chart metadata in Arabic presents unique challenges at the intersection of cross-lingual transfer and data-to-text generation. Chart-to-text benchmarks have advanced English-language research, yet Arabic remains without a comparable resource, underscoring its continued underrepresentation in NLP. To cover this gap, we construct the first Arabic ChartSumm benchmark by translating chart metadata and reference summaries from English into Modern Standard Arabic (MSA). Two high-quality machine translation models with contrasting architectures are employed: NLLB-200-distilled-600M, designed for low-resource coverage, and Qwen2.5-1.5B, an open large language model with general multilingual capabilities. A central contribution of this work is a translation quality evaluation that systematically assesses both systems using BLEU, chrF, COMET_ref, and COMET_QE metrics against a Google-Translate Arabic pivot. Results demonstrate that NLLB achieves markedly higher lexical and semantic fidelity. Building on this foundation, we fine-tune two models, mT5 (multilingual) and CAMeL-Lab's AraBART (Arabic-specific), to generate Arabic summaries from structured chart metadata. Experimental results show that AraBART trained on NLLB translations outperforms other configurations, achieving ROUGE-L = 63.8 and BLEU = 33.1, highlighting the strong dependency of downstream summarization quality on translation accuracy and demonstrating its superior capacity for Arabic generation.

Keywords: Chart-to-text generation, Arabic NLP, Arabic summarization, Machine translation

1. Introduction

Chart-to-text generation is the task of producing natural language summaries from structured data such as tables, graphs, and charts. It plays a key role in making data more accessible, especially in domains like journalism, education, and assistive technology. While recent progress has been driven by large-scale benchmarks such as ChartSumm in English (Rahman et al., 2023), no equivalent dataset exists for Arabic, one of the world's most widely spoken yet underrepresented languages in Natural Language Processing (NLP).

Arabic poses unique challenges for natural language generation due to its rich morphology, flexible syntax, and limited availability of annotated resources (Mashaabi et al., 2024). These factors make it difficult to train and evaluate models for Arabic summarization, especially in structured data scenarios like chart description. Moreover, most pre-trained models are optimized for high-resource languages, and their cross-lingual transfer to Arabic is not well understood in data-to-text settings.

To address this gap, we create an Arabic version of the ChartSumm benchmark by translating chart metadata and summaries into MSA. We use two strong machine translation models, Meta's NLLB-200-distilled-600M (NLLB Team et al., 2022) and Alibaba's Qwen2.5-1.5B (Yang et al., 2024) to investigate how translation quality affects downstream summarization. Our aim is to study how well multilingual and Arabic-specific models can generate accurate and fluent Arabic summaries from structured chart data.

We fine-tune two pre-trained models, mT5 (Xue et al., 2021) and CAMeL-Lab's AraBART (Kamal Eddine et al., 2022), selected for their strong performance in Arabic summarization tasks (Masri et al., 2025). Fine-tuning is performed on the translated ChartSumm dataset (Figure 1), where inputs consist of flattened chart metadata (titles, axis labels, and data series) formatted into structured natural language prompts. Models' outputs are evaluated using standard metrics (BLEU, ROUGE-1/2/L/Lsum) and complementary measures (BLEURT, Perplexity, CIDEr, and Content Selection) to assess fluency, informativeness, and faithfulness to the reference text. This paper makes the following contributions:

- We create and release a translated version of ChartSumm in Arabic, the first dataset of its kind for Arabic chart-to-text generation.
- We compare two MT-based translation strategies and show their impact on the quality of Arabic summaries.
- We benchmark mT5 and AraBART on this task and provide guidance for improving multilingual summarization in low-resource settings.

To the best of our knowledge, this work presents the first Arabic extension of the ChartSumm bench-

[*] The authors contributed equally to this work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10447–10456
11–16 May 2026.

mark, supporting the generation of contextualized natural language summaries from structured chart metadata in Arabic MSA. The rest of this paper is organized as follows: Section 2 reviews related work spanning both Arabic and other languages, while Section 3 describes the dataset used in our study, which was translated into Arabic using two different translation models and the evaluation metrics adopted to evaluate the translation quality. Section 4 outlines the experimental setup, beginning with the use of the BART-Large-CNN as a base model, which achieved the highest BLEU score on the "test-s" subset reported by (Rahman et al., 2023), followed by the fine-tuning process for Arabic chart summarization dataset, and concluding with a discussion of the summarization model's results. Finally, Section 5 concludes the paper by summarizing the key findings and discussing potential directions for future research.

2. Related Work

Although there is substantial work on Arabic text summarization spanning both abstractive and extractive methods with pretrained models such as AraT5 (Nagoudi et al., 2022) and AraBART (Eddine et al., 2022), alongside deep learning approaches like pointer-generator LSTMs (Al-Maleh and Desouki, 2020) and encoder–decoder models incorporating NER and copy mechanisms trained on headline datasets (Essa et al., 2025), none of these studies explicitly explore generating Arabic summaries from chart metadata (e.g., axis labels, legends, or numeric values). Related research has also investigated semantic-graph-based methods (e.g., SemG-TS) (Etaiwi and Awajan, 2022), BERT-based fine-tuning (Abdelwahab et al., 2023; Elmadani et al., 2020), and earlier LSA-based extractive techniques (Al Qassem et al., 2017); however, all of these focus on summarizing free-form Arabic text rather than structured inputs from charts, making Arabic chart-to-text summarization a largely unexplored research area.

In contrast to the lack of Arabic work, there is significant research in other languages—especially English—focused on generating summaries directly from charts. For example, the Chart-to-Text benchmark (Kantharaj et al., 2022) introduced a large-scale dataset of over 44,000 charts, revealing that although models can generate fluent captions, they often hallucinate trends or misinterpret complex patterns. In addition, the ChartSumm benchmark proposed by (Rahman et al., 2023), introduced a large-scale dataset of 84,363 charts covering diverse topics and chart types, each annotated with both short and long natural language summaries. While the original ChartSumm dataset is in English, the authors also explored the possibility of extending it to other languages using automated translation tools. Follow-up studies include models such as ChartAdapter (Xu et al., 2024), a vision-language model trained on roughly 190k chart-summary pairs that outperforms earlier table-to-text baselines, and ChartThinker (Liu et al., 2024), which incorporates chain-of-thought reasoning to enhance logical coherence. UniChart (Masry et al., 2023) further advanced the field by pretraining on chart visuals, data, and text, achieving state-of-the-art results for summarization and reasoning, while earlier approaches like the Chart-to-Text Transformer (Obeid and Hoque, 2020) relied on template-based and planning strategies. In addition, surveys such as Natural Language Generation for Visualizations (Hoque and Islam, 2025) have provided a comprehensive overview of methods ranging from CNN+LSTM pipelines to transformer-based systems, noting common error types like "value" and "trend" inaccuracies and comparing prompting versus fine-tuning strategies. In scientific contexts, models that combine ResNet, OCR, and LSTM (Tan et al., 2022) have shown improved BLEU scores when summarizing research charts. Altogether, this body of work highlights that while chart-to-text summarization has gained traction in English and other languages, it remains largely unexplored within Arabic NLP.

3. The ChartSum Dataset

This section outlines the dataset used in our study, the English-to-Arabic translation process, and the evaluation approach adopted to assess translation quality.

3.1. Dataset Construction

The dataset used in this study is derived from the **ChartSumm** benchmark (Rahman et al., 2023), a large-scale English-language resource for automatic chart summarization. ChartSumm consists of 84,363 chart–summary pairs aligned with chart images and structured metadata, providing both short and long-form summaries. It integrates data from two major sources: *Knoema* (43,179 examples), which provides concise, factual summaries generated by the Yodatai assistant, and *Statista* (41,184 examples), which contains longer, human-authored summaries. In this study, we use only the Statista subset of ChartSumm.This subset was selected because it offers rich, descriptive annotations and greater structural diversity, making it better suited for evaluating downstream summarization tasks in Arabic. As described by (Rahman et al., 2023), Statista is an online platform that publishes statistical information on a wide range of topics, each accompanied by a short human-written description.

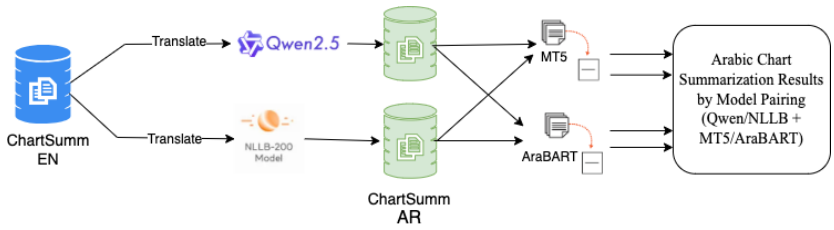

Figure 1: Arabic Chart Summarization pipeline: English ChartSumm is translated into Arabic using Qwen2.5-1.5B and NLLB-200-distilled-600M, then summarized with mT5 and AraBART.

The topics span multiple domains, including economics, marketing, industry, and opinion research.

For dataset construction, (Rahman et al., 2023) crawled over 750,000 publicly available Statista pages to collect 41,184 chart–summary pairs, each containing chart metadata, textual summaries, and associated chart images. The data were then categorized into *simple* and *complex* charts based on the number of columns in each chart. Similar to the Knoema subset, tokenization and stemming were applied to normalize text data. Since many examples lacked explicit `x_label` values, heuristic rules were introduced to automatically classify the `x_labels` as *Year*, *Month*, *Day*, *Quarter*, *Country*, *City*, or *Continent*.

Each instance in our selected subset includes a structured JSON representation of a chart comprising the chart title, axis labels (`x_label` and `y_label`), and tabular data mapping categories to their corresponding numerical values, as well as a human-authored English summary and a filename linking to the associated chart image (see Listing 1).

To adapt this benchmark for Arabic, we translated all metadata fields and reference summaries into MSA. Translation was conducted using two high-quality machine translation systems as reported in (Alrashed et al., 2024): Meta's NLLB-200-distilled-600M and Alibaba's Qwen2.5-1.5B. The Arabic dataset preserves the original structure, with translated chart titles, axis labels, and descriptions, while keeping the numerical values unchanged. An example instance of an English chart and Arabic equivalent is shown in Table 1.

Listing 1: English Chart metadata JSON example

```
{
    "x_label": "x_label",
    "y_label": [
      "Which party would you vote for
          in the May 2019 European
        Parliament elections?"
    ],
    "data": {
      "x_label": [
        "The Brexit Party",
        "Liberal Democrats",
        "Labour",
        "Green",
        "Conservatives",
        "Other",
        "Change UK",
        "UKIP"
      ],
      "Which party would you vote for
          in the May 2019 European
        Parliament elections?": [
        37.0,
        19.0,
        13.0,
        12.0,
        7.0,
        6.0,
        4.0,
        3.0
      ]
    },
    "title": "UK voting intention in
        the European Parliament
        elections as of May 2019",
    "summary": "This statistic
        presents the voting intention
         of adults in the United
        Kingdom , for the European
        Elections due to take place
        on May 23 , 2019 . The
        recently formed Brexit Party
        had the highest share of
        adults intending to vote for
        them at 37 percent , with the
         governing Conservative Party
         trailing in fifth at just
        seven percent .",
    "image": "train_s_0.png"
}
```

This bilingual setup enables a controlled evaluation of how translation quality influences downstream chart-to-text generation in Arabic. The resulting Arabic corpus comprises 41,184 chart–summary pairs, maintaining the same data splits as the original Statista subset of ChartSumm: **32,985** pairs for training, **4,101** for validation, and **4,098** for testing. Each split contains chart metadata and human-authored Arabic summaries translated from the original English descriptions.

Chart Meta-data	Original English Text	Arabic Translation (NLLB-200-distilled-600M)	Arabic Translation (Qwen 2.5-1.5B)
Title	UK voting intention in the European Parliament elections as of May 2019	نية التصويت في المملكة المتحدة في انتخابات البرلمان الأوروبي اعتبارًا من مايو ٢٠١٩	التصويت في الجمعية الأوروبية للبرلمان البريطاني في الوقت الحاضر
x_label	x_label	x_label	x_label
y_label	Which party would you vote for in the May 2019 European Parliament elections?	إلى أي حزب ستصوت في انتخابات المجلس الأوروبي في مايو ٢٠١٩؟	هل ستختار أي حزب في انتخابات مجلس الأمة الأوروبي في مايو ٢٠١٩؟
data [x_label]	The Brexit Party, Liberal Democrats, Labour, Green, Conservatives, Other, Change UK, UKIP	حزب خروج بريطانيا، الديمقراطيين الليبراليين، العمل، الخضراء، المحافظون، غيرها، تغيير المملكة المتحدة (UKIP)	['الحزب البريطاني للبريكست', 'الخضر', 'العمال', 'الليبراليين', 'غير ذلك', 'اتيفون conservال', 'UKIP', '
data [y_label]	37.0, 19.0, 13.0, 12.0, 7.0, 6.0, 4.0, 3.0	٠,٧٣، ٠,٣١، ٠,٩١، ٠,٢١، ٠,٧، ٠,٦، ٠,٤، ٠,٣	37.0, 19.0, 13.0, 12.0, 7.0, 6.0, 4.0, 3.0
Summary	This statistic presents the voting intention of adults in the United Kingdom , for the European Elections due to take place on May 23 , 2019 . The recently formed Brexit Party had the highest share of adults intending to vote for them at 37 percent , with the governing Conservative Party trailing in fifth at just seven percent.	هذه الإحصاءات تعرض نية التصويت من البالغين في المملكة المتحدة ، في الانتخابات الأوروبية التي من المقرر أن تجري في ٢٣ مايو ٢٠١٩ . كان لدى حزب خروج بريطانيا من الاتحاد الأوروبي الذي تم تشكيله حديثًا أعلى نسبة من البالغين الذين يعتزمون التصويت لهم بنسبة ٣٧ في المائة ، مع تراجع الحزب المحافظ الحاكم في المركز الخامس بنسبة سبعة في المائة فقط.	هذا التقرير يعرض توقعات الناخبين الأكبر سنا في المملكة المتحدة للاستفتاء الأوروبي الذي سيقام يوم ٢٣ مايو ٢٠١٩.كانت حزب بريكست الجديد الذي تم تشكيله أقوى ممثلة من الناخبين الذين كانوا على استعداد للتصويت له بنسبة ٣٧%، بينما جاء الحزب الديمقراطي المتحكم في الحكومة في المرتبة الخامسة بنسبة ٧
Image	train_s_0.png	train_s_0.png	train_s_0.png

t

Table 1: Example of an English ChartSumm instance and its Arabic counterpart, where both metadata and summaries are translated into Arabic MSA using NLLB-200-distilled-600M and Qwen2.5-1.5B-Instruct.

3.2. Translation Quality Evaluation

We quantitatively evaluate Arabic translation quality and directly compare two MT systems NLLB and Qwen built from the same English metadata. Using a Google-Translate Arabic pivot as a proxy reference, we compute four complementary metrics over the merged test set: BLEU (1–4g, smoothed) (Papineni et al., 2002), chrF (Popovic, 2015) (n=6, β=2), COMET_ref (reference-based) (Rei et al., 2020), and COMET_QE (Rei et al., 2022) ("quality estimation", reference-less). Results in Table 2 show that NLLB outperforms Qwen across all metrics by large margins ($+13.53$ BLEU, $+20.39$ chrF, $+0.087$ COMET_ref, $+0.245$ COMET_QE). We therefore identify NLLB as the stronger Arabic translation source for our task and retain Qwen for comparative and error analysis. A pivot reference can bias style and lexical overlap; we mitigate this by reporting both overlap (BLEU/chrF) and learned semantic estimators (COMET), by releasing per-example COMET scores for targeted inspection.

4. Experiments

We evaluate the effectiveness of Arabic chart metadata summarization through two consecutive downstream tasks: machine translation and Arabic summary generation, as described in Figure 1.

4.1. Evaluation Metrics

In addition to standard metrics such as BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, we adopt the same evaluation measures

System	BLEU	chrF	COMET_ref	COMET_QE
NLLB	**29.65**	**62.24**	**83.9**	**32.8**
Qwen	16.12	41.848	75.20	8.30

Table 2: Automatic translation quality comparison against a Google-Translate Arabic pivot. Higher is better.

employed by (Rahman et al., 2023) to provide a comprehensive assessment of summary quality. These metrics include: *(1) BLEURT*, which evaluates the fluency of the generated text and its semantic alignment with the reference; *(2) Perplexity (PPL)*, which measures how well a probabilistic language model predicts a given text sequence. In this work, we employ the 'aubmindlab/aragpt2-base' model instead of the standard 'pre-trained GPT-2', as our evaluation is conducted on Arabic data; *(3) CIDEr*, which measures the average cosine similarity of n-gram representations between the candidate and reference summaries, thereby capturing grammaticality and semantic richness; and *(4) Content Selection (CS)*, which assesses how closely the generated summary reflects the information contained in the reference text.

4.2. Baseline

For the baseline, we employ the BART-Large-CNN model, which achieved the highest BLEU score on the "test-s" subset reported by (Rahman et al., 2023). According to their findings, BART was selected due to its demonstrated effectiveness in chart-to-text generation tasks. Although they fine-tuned BART-Large-CNN on the ChartSumm dataset, the fine-tuned model is not publicly available in their GitHub repository. Owing to limited computational resources, we therefore adopt the publicly released pretrained BART-Large-CNN model without additional fine-tuning as our baseline. We adopt the same experimental configuration, training the BART-Large-CNN model for three epochs with a batch size of eight. The fine-tuning process employs an initial learning rate of 1e 6, using the AdamW optimizer and cross-entropy as the loss function.

To establish a cross-lingual baseline, we implemented a pivot summarization approach in which the English BART-Large-CNN summarization model operates directly on the chart metadata. It was used to generate English summaries from the original chart metadata of Rahman et al.. The resulting English summaries were then translated into MSA using NLLB-200-distilled-600M and Qwen2.5-1.5B. This pipeline (*metadata → English summarization → Arabic translation*) serves as a pivot baseline for evaluating Arabic chart-to-text generation through cross-lingual transfer.

In addition, the *test-s* subset of the ChartSumm dataset was translated into Arabic using the Google Cloud Translation API (v2), accessed through the official `google-cloud-translate` Python client library. This API interfaces with Google's Neural Machine Translation (NMT) system, which powers the public Google Translate service. Furthermore, a subset of approximately 1,000 Arabic translated chart–summary pairs was manually reviewed and post-edited to ensure translation quality and to enable a fair comparison with the outputs generated by the NLLB and Qwen models.

Tables 3 and 4 present the evaluation results of the pivot-based Arabic summarization baseline using lexical, semantic, and fluency-based metrics. Across all measures, the NLLB model consistently outperforms Qwen, confirming its effectiveness as a translation component within the summarization pipeline. As shown in Table 3, NLLB achieves higher BLEU (14.08) and ROUGE scores across all variants, indicating stronger lexical alignment and better content preservation with respect to human references. Complementary results in Table 4 further highlight NLLB's superiority in semantic fidelity and fluency, with notably higher BLEURT (22.03), CIDEr (60.90), and Content Selection (85.90) scores. as well as a much lower Perplexity (78.42). The relative percentages of perplexity results reveal a pronounced disparity in fluency: NLLB's value of 78.42 indicates relatively coherent and well-structured text generation, whereas Qwen's perplexity is approximately 883.7 higher than that of human references, reflecting substantial degradation in linguistic fluency and syntactic consistency. Overall, these results demonstrate that NLLB produces more accurate, semantically faithful, and linguistically fluent Arabic summaries, establishing it as the stronger pivot translation model for downstream fine-tuning and evaluation.

4.3. Fine-tuning Process

As mentioned earlier (Secotion 3), all chart metadata and reference summaries were translated into MSA using two machine translation models: NLLB-200-distilled-600M and Qwen2.5-1.5B-Instruct. Unlike the pivot baseline, which translates only English-generated summaries, this stage covers the full dataset, including chart titles, axis labels, and human-authored summaries.

To generate Arabic summaries directly from

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
NLLB	**14.08**	**37.10**	**21.72**	**34.73**	**34.74**
Qwen	9.76	35.15	19.77	32.73	32.70

Table 3: Evaluation of pivot-based Arabic summaries using BLEU and ROUGE variants. Higher is better.

Model	BLEURT	CIDEr	CS	PPL
NLLB	**22.03**	**60.90**	**85.90**	**78.42**
Qwen	13.60	37.06	78.06	883.7

Table 4: Semantic similarity, content selection, and fluency evaluation of pivot baseline Arabic summaries.

chart metadata, we fine-tune two models: mT5 (`google/mt5-base`) (Xue et al., 2020), a multilingual transformer trained on 101 languages, and AraBART (`moussaKam/AraBART`) (Kamal Eddine et al., 2022), a BART-based model pre-trained on Arabic corpora. Fine-tuning is performed on the translated ChartSumm dataset using the same official splits as the original English corpus. Each input instance is constructed by linearizing chart metadata, including titles, axis labels, and values, into structured Arabic prompts.Table 5 presents examples of chart summarization outputs generated by the fine-tuned mT5 and AraBART models using inputs translated by NLLB and Qwen.

We use a maximum input length of 512 tokens and a target summary length of 128 tokens. The models are optimized with AdamW (lr: 5e-5), a batch size of 4, and up to 5 epochs, with early stopping based on validation ROUGE-L score.

4.4. Results

Across both training and test evaluations (Table 6 and Table 7), the NLLB+AraBART configuration demonstrates clear superiority over all other model pairings. During fine-tuning, it attains the highest validation ROUGE-L score of 57.6, outperforming NLLB+mT5 (53.5), Qwen+mT5 (29.0), and Qwen+AraBART (31.2). On the test set, NLLB+AraBART achieves the best overall results across all metrics, including ROUGE-1 = 66.3, ROUGE-2 = 42.2, ROUGE-L = 63.8, ROUGE-Lsum = 63.8, and BLEU = 33.1. In contrast, the Qwen+AraBART model performs the weakest, with ROUGE-L of only 22.6 and BLEU of 13.0, reflecting the negative impact of translation noise. These findings confirm that pairing the high-quality NLLB translations with an Arabic-optimized model (AraBART) yields the most effective Arabic summarization results in terms of both lexical overlap and content fidelity.

Furthermore, we utilize the manually reviewed set of 1,000 translated chart–summary pairs previously employed to evaluate the baseline model. To ensure the reliability of this subset, we conducted a focused manual post-editing and quality-control process on these instances, all of which were drawn from the original ChartSumm English test set. The summaries were translated into Modern Standard Arabic using the NLLB+AraBART and subsequently reviewed by two native Arabic-speaking researchers with expertise in computational linguistics and machine translation. The annotators were instructed to verify (i) semantic faithfulness to the English source, (ii) correctness of numerical values and chart-specific terminology, (iii) grammatical well-formedness, and (iv) naturalness and stylistic adequacy in Modern Standard Arabic. Corrections were applied when necessary, including lexical refinements, syntactic restructuring, normalization of numerical expressions, and terminology standardization. The manually reviewed subset is explicitly identified in the dataset release and serves as a higher-confidence evaluation partition to support future research requiring quality-controlled Arabic references.

Tables 8 and 9 present the performance of the fine-tuned models across both standard evaluation metrics and complementary assessment measures, using this manually verified subset. When compared to the pivot-based baselines in Tables 3 and 4, the fine-tuned models show substantial improvements in both lexical and semantic quality.

Notably, the NLLB+AraBART configuration achieves the strongest overall results, surpassing the baseline NLLB model across all evaluation dimensions. BLEURT increases from 22.03 to 23.75, while CIDEr rises markedly from 60.90 to 86.24, reflecting enhanced semantic alignment and greater content fidelity between generated and reference summaries. The Content Selection (CS) score also improves slightly (from 85.90 to 86.33), indicating stronger information retention from the source chart metadata. In terms of fluency, measured by Perplexity (PPL), the fine-tuned models demonstrate a clear improvement over the pivot-based baselines. The NLLB+AraBART model achieves the highest fluency with a relative perplexity of 85.13, outperforming both the multilingual NLLB+mT5 (63.47) and the baseline NLLB model (78.42). This improvement confirms that fine-tuning AraBART on NLLB-translated data yields smoother, more coher-

		Chart meta data ID: Test_s_819
Chart title		Forecast: value of shipments precision turned product manufacture US 2008-2020
English Chart Summary		This forecast statistic shows the value of shipments of precision turned product manufacture in the United States from 2008 to 2013 , with forecasts up until 2020 . By 2016 , value of shipments of precision turned product manufacture in the United States are projected to reach approximately 21.41 billion U.S. dollars.
Arabic Translated NLLB Version	Gold	تظهر هذه الإحصاءات المتوقعة قيمة شحنات تصنيع المنتجات المتحولة بدقة في الولايات المتحدة من ٢٠٠٨ إلى ٢٠١٣ ، مع توقعات حتى ٢٠٢٠. بحلول عام ٢٠١٦ ، من المتوقع أن يصل قيمة شحنات تصنيع المنتجات المتحولة بدقة في الولايات المتحدة إلى حوالي ٤١.٢١ مليار دولار أمريكي .
	mT5	تظهر هذه الإحصاءات المتوقعة قيمة الشحنات من تصنيع المنتجات المتحولة بدقة في الولايات المتحدة من ٢٠٠٨ إلى ٢٠١٣ مع توقعات حتى ٢٠٢٠ .بحلول عام ٢٠١٦ ، من المتوقع أن تصل قيمة شحنات تحويل دقة المنتجات التحويلية بدقة إلى حوالي ٢٣ مليار دولار أمريكي .
	AraBART	تظهر هذه الإحصاءات المتوقعة قيمة شحنات تصنيع المنتجات المتحولة بدقة في الولايات المتحدة من ٢٠٠٨ إلى ٢٠١٣ ، مع توقعات حتى ٢٠٢٠. بحلول عام ٢٠١٦ ، من المتوقع أن يصل قيمة شحنات منتجات تصنيع المنتجات المتحولة بدقة إلى حوالي ٤٠.٢١ مليار دولار أمريكي .
Arabic Translated Qwen Version	Gold	هذا التقرير الإحصائي يظهر قيمة الشحنات للمنتجات المصنعة بعمق تم توليدها في الولايات المتحدة من عام ٢٠٠٨ إلى عام ٢٠١٣ ، مع توقعات حتى عام ٢٠٢٠. حوالي ٤١.٢١ مليار دولار أمريكي في عام ٢٠١٦.
	mT5	هذا التقرير الإحصائي يظهر قيمة الشحنات لصناعة المعدات الصناعية في الولايات المتحدة من عام ٢٠٠٨ إلى عام ٢٠١٣ ، مع توقعات حتى عام ٢٠٢٠. حوالي ٢٧.١ مليار دولار أمريكي في عام ٢٠١٦.
	AraBART	تظهر هذه الإحصاءات المتوقعة قيمة شحنات تصنيع المنتجات المتحولة في الولايات المتحدة من ٢٠٠٨، مع توقعات حتى ٢٠٢٠. بحلول عام ٢٠١٦ ، من المتوقع أن يصل قيمة شحنات منتجات تصنيع المنتجات المتحولة إلى حوالي ٤٠.٢١ مليار دولار أمريكي .

Table 5: Chart summarization outputs from the fine-tuned models based on mT5 and AraBART

Model	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5	Best ROUGE-L
NLLB+mT5	38.3	39.3	49.9	53.5	53.5	53.5
NLLB+AraBART	54.9	56.0	56.8	57.2	57.6	**57.6**
Qwen+mT5	27.5	28.0	28.3	28.2	29.0	29.0
Qwen+AraBART	28.1	28.3	29.1	28.3	31.0	31.2

Table 6: ROUGE-L scores across epochs for different models at **validation data**.

ent, and linguistically consistent Arabic summaries.

Conversely, the Qwen-based models continue to lag across all metrics. The Qwen+mT5 combination achieves moderate results (ROUGE-L = 31.41, BLEU = 9.07), while Qwen+AraBART performs substantially worse (ROUGE-L = 14.95, BLEU = 0.65), suggesting that translation noise and weaker semantic fidelity in Qwen outputs negatively affect

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BLEU
NLLB+mT5	65.7	38.2	63.2	63.1	21.5
NLLB+AraBART	**66.3**	**42.2**	**63.8**	**63.8**	**33.1**
Qwen+mT5	60.2	29.9	57.8	57.8	17.3
Qwen+AraBART	23.5	7.5	22.6	22.6	13.0

Table 7: Evaluation metrics (ROUGE and BLEU) for different summarization models at the **Test data**.

summarization quality. In terms of fluency, the Qwen+mT5 model attains a moderate perplexity score (56.71), whereas Qwen+AraBART records the lowest score (20.16), indicating disfluent and unstable sentence generation. Nonetheless, both Qwen-based models exhibit notable improvements relative to the original Qwen baseline, whose perplexity was approximately 883.7 higher than that of human references.

Overall, these findings confirm that fine-tuning Arabic-specific models (AraBART) on high-quality NLLB translations yields significant improvements over the pivot-based baselines in both informativeness and fluency, establishing NLLB+AraBART as the most effective configuration for Arabic chart summarization.

5. Conclusion & Future Work

This study presents the first Arabic extension of the ChartSumm benchmark, enabling contextual Arabic summary generation from structured chart metadata. The English ChartSumm dataset was translated into Modern Standard Arabic (MSA) using two machine translation models, Qwen2.5-1.5B and NLLB-200-distilled-600M, with a manually reviewed subset of 1,000 test pairs used to assess translation quality, confirming the superior performance of NLLB. The fine-tuned mT5 (multilingual) and AraBART (Arabic-specific) models on the translated corpus demonstrate that translation quality strongly influences summarization performance, with the NLLB+AraBART combination outperforming all baselines across BLEU, ROUGE, BLEURT, and CIDEr metrics.

These findings underscore the importance of preserving semantic and numerical fidelity in Arabic summarization and structuring input formats to retain contextual meaning. Future work will focus on expanding human-verified translations, conducting human evaluations of summary quality, and incorporating visual chart features to enhance data-to-text generation in Arabic.

6. Dataset Release and Availability

Arabic ChartSumm is publicly released to support research on Arabic chart-to-text generation.[1] The released resource includes the full Arabic translation of the ChartSumm dataset generated using NLLB-200-distilled-600M, which demonstrated the highest translation quality in our evaluation. The dataset preserves the original ChartSumm train, validation, and test splits to ensure comparability with prior work.

7. Limitations

While Arabic ChartSumm establishes the first benchmark for Arabic chart-to-text summarization, several limitations remain.

First, the dataset is derived through machine translation rather than being originally authored in Arabic. Although we employ multiple automatic evaluation metrics and a manually reviewed subset to validate quality, translated references may not fully capture the stylistic richness and discourse conventions of naturally written Arabic summaries.

Second, translation quality is partially evaluated using a pivot reference generated by Google Translate. While this enables consistent lexical comparison, it may introduce stylistic bias toward that reference. We mitigate this through COMET-QE and manual inspection; however, fully human-authored Arabic references would provide a stronger evaluation standard.

Third, our evaluation relies primarily on automatic summarization metrics (ROUGE and BLEU). Human evaluation of generated summaries would provide deeper insights into luency, faithfulness, and readability, and is planned for future work. Finally, although we compare two strong translation systems, a broader comparison including additional MT models could further contextualize translation robustness.

Despite these limitations, we believe Arabic ChartSumm provides a valuable foundation for advancing Arabic-native data-to-text generation and enables future extensions involving fully human-authored resources and human-centered evaluation protocols.

[1] https://github.com/a-fashwan/Arabic-ChartSumm-Dataset.git

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
NLLB+mT5	36.72	20.43	34.96	34.95
NLLB+AraBART	**39.45**	**24.26**	**37.50**	**37.54**
Qwen+mT5	32.63	16.57	31.41	31.40
Qwen+AraBART	15.40	08.69	14.95	14.93

Table 8: ROUGE evaluation results for pivot-based Arabic summarization models using NLLB and Qwen translation combined with mT5 and AraBART summarizers.

Model	BLEU	BLEURT	CIDEr	CS	PPL
NLLB+mT5	11.09	20.83	56.97	83.28	63.47
NLLB+AraBART	**15.36**	**23.75**	**86.24**	**86.33**	**85.13**
Qwen+mT5	9.07	16.85	35.76	73.42	56.71
Qwen+AraBART	0.65	10.48	17.67	41.24	20.16

Table 9: Evaluation results for Arabic summarization models using different translation–summarization combinations, reported across BLEU, BLEURT, CIDEr, Content Selection (CS), and Perplexity (PPL) metrics.

8. References

Mohamed Yassin Abdelwahab, Yazeed Al Moaiad, and Zainab Binti Abu Bakar. 2023. Arabic text summarization using pre-processing methodologies and techniques. *Asia-Pacific Journal of Information Technology & Multimedia*, 12(1).

Molham Al-Maleh and Said Desouki. 2020. Arabic text summarization using deep learning approach. *Journal of Big Data*, 7(1):109.

Lamees Mahmoud Al Qassem, Di Wang, Zaid Al Mahmoud, Hassan Barada, Ahmad Al-Rubaie, and Nawaf I Almoosa. 2017. Automatic arabic summarization: a survey of methodologies and systems. *Procedia Computer Science*, 117:10–18.

Sultan Alrashed, Dmitrii Khizbullin, and David R. Pugh. 2024. FineWeb-Edu-Ar: Machine-translated Corpus to Support Arabic Small Language Models. Technical Report arXiv:2411.06402, King Abdullah University of Science and Technology (KAUST). CC-BY-NC-4.0 licensed; contains 202B tokens.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization. *arXiv preprint arXiv:2203.10945*.

Khalid N Elmadani, Mukhtar Elgezouli, and Anas Showk. 2020. Bert fine-tuning for arabic text summarization. *arXiv preprint arXiv:2004.14135*.

Nada Essa, MM El-Gayar, and Eman M El-Daydamony. 2025. Enhanced model for abstractive arabic text summarization using natural language generation and named entity recognition. *Neural Computing and Applications*, pages 1–23.

Wael Etaiwi and Arafat Awajan. 2022. Semg-ts: Abstractive arabic text summarization using semantic graph embedding. *Mathematics*, 10(18):3225.

Enamul Hoque and M Saidul Islam. 2025. Natural language generation for visualizations: State of the art, challenges and future directions. In *Computer Graphics Forum*, volume 44, page e15266. Wiley Online Library.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Mengsha Liu, Daoyuan Chen, Yaliang Li, Guian Fang, and Ying Shen. 2024. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. *arXiv preprint arXiv:2403.11236*.

Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A survey of large language models for arabic language and its dialects. *arXiv preprint arXiv:2410.20238*.

Sari Masri, Yaqeen Raddad, Fidaa Khandaqji, Huthaifa I. Ashqar, and Mohammed Elhenawy.

2025. Transformer models in education: Summarizing science textbooks with arabart, mt5, arat5, and mbart. In *Intelligent Systems, Blockchain, and Communication Technologies*, volume 1268 of *Lecture Notes in Networks and Systems*, pages 286–300. Springer, Cham.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*. Describes the NLLB-200 model family, including the distilled 600M variant.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*. Version v3, last revised 11 Jun 2023.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hao Tan, Chen-Tse Tsai, Yujie He, and Mohit Bansal. 2022. Scientific chart summarization: Datasets and improved text modeling. In *SDU@AAAI*.

Peixin Xu, Yujuan Ding, and Wenqi Fan. 2024. Chartadapter: Large vision-language model for chart summarization. *arXiv preprint arXiv:2412.20715*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 483–498.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*. Describes the Qwen2.5 series including the 1.5B variant.