

Phrase-Grounding-Aware Supervised Fine-Tuning for Chart Recognition via Side-Masked Attention

Supplementary Material

This supplementary document provides additional details that complement the main paper, including information that could not be presented in the main manuscript. Specifically, we provide:

- A supplementary explanation of how the logit-contribution formulation of Ferrando et al. [3], which was originally introduced as an inference-time interpretability method, is reformulated and incorporated into our training framework.
- Additional implementation details for our method.
- Additional results and visualizations.
- Dataset details and visual examples, including the full GPT prompts used in our data-generation pipeline and representative generated samples, together with a brief description of the separate annotation protocol used in C2T.
- Limitations of our method.

A1. Adapting the logit-contribution formulation to our method

First, we provide a brief overview of the logit-contribution formulation introduced by Ferrando et al. [3]. The output probability distribution of an L -layer transformer language model is computed by applying a softmax to the product of the final-layer hidden state $x_t^L \in \mathbb{R}^d$ and the output embedding matrix $U \in \mathbb{R}^{d \times |\mathcal{V}|}$: $P(x_t^L) = \text{softmax}(x_t^L U)$, where $|\mathcal{V}|$ denotes the size of the vocabulary. Since each transformer layer employs residual connections, the final hidden state can be decomposed as

$$x_t^L = \sum_{l=1}^L o_t^l + \sum_{l=1}^L z_t^l + x_t^0, \quad (10)$$

where o_t^l and z_t^l are, respectively, the FFN and self-attention outputs of layer l , and x_t^0 is the input embedding. The unnormalized logit for the output token w_t is the inner product of x_t^L with the w_t -th column of U , denoted U_{w_t} :

$$\begin{aligned} \text{logit}_{w_t} &= x_t^L U_{w_t} \\ &= \left(\sum_{l=1}^L o_t^l + \sum_{l=1}^L z_t^l + x_t^0 \right) U_{w_t} \end{aligned} \quad (11)$$

$$= \sum_{l=1}^L (o_t^l U_{w_t}) + \sum_{l=1}^L (z_t^l U_{w_t}) + x_t^0 U_{w_t}. \quad (12)$$

Ferrando et al. focus on attributing contributions to the self-attention outputs, since past tokens influence the prediction

through attention. Denoting the j -th input to layer l by $x_j^{l-1} \in \mathbb{R}^d$, the self-attention output at layer l and position t is

$$z_t^l = \sum_{j=1}^t T_{t,j}^l(x_j^{l-1}), \quad (13)$$

where we omit bias terms for clarity and the term $T_{t,j}^l(\cdot)$ in Eq. (13) corresponds to Eq. (2) in the main paper.

We incorporate the LayerNorm L_N into Eq. (1) to match the architecture of our VLMs. By accounting for the attention outputs $z_t^{l,+}$, the FFN outputs $\text{FFN}^l(\cdot)$ across all layers, and the input embedding x_t^0 , we ensure that the SMAM logit computed for $\pi_{w_t}^+$ (Eq. (7) in the main paper) remains aligned with the full logit decomposition in Eq. (12), while still being computed under the masked-attention constraint imposed by the mask M_t . In contrast to Ferrando et al., who decompose attention and FFN contributions and analyze them separately, our formulation integrates these components at each layer and accumulates them across layers following the residual computation.

In standard attention computation, the weighted sum over value vectors is performed in a single aggregation step, so only the hidden states for the current output token need to be stored. In contrast, computing the per-patch contribution $\Delta \text{logit}_{w_t \leftarrow x_j}$ as in Eq. (3) requires retaining a separate vector for every vision patch token at each $x_t \in \mathcal{X}$, resulting in an additional memory cost proportional to $(\text{number of vision patches}) \times (\text{hidden dimension}) \times |\mathcal{X}|$, where \mathcal{X} is the set of phrase-mask aligned tokens. This makes direct per-patch computation prohibitively expensive, especially during training. SMAM avoids this cost by aggregating contributions at the mask-region level rather than at the per-patch level. During training, we only need to compute $\pi_{w_t}^+$ for each x_t , which introduces an additional memory cost of approximately $(\text{hidden dimension}) \times |\mathcal{X}|$, making it a much more practical requirement.

To make this computation feasible, we leverage the formulations in Eqs. (4), (6), and (7) which enable this region-level aggregation in a single forward pass. As a result, SMAM eliminates the need to store per-patch contributions, and the overall computational overhead scales linearly with the number of phrase-aligned tokens in each sample.

Eq. (7) accumulates the SMAM outputs $z_t^{l,+}$ and the corresponding FFN outputs at each layer l , computed from the layer input $X^{l-1} = (x_1^{l-1}, \dots, x_t^{l-1})$, across layers following the residual connections of the transformer. However, the layer input X^{l-1} is inherited from the original trans-

Table 6. Hyperparameters for each model across tasks. “Bz” indicates the global batch size and “Epoch” is the number of training epochs.

Model	C2T		ChartQA		QA-CoT		GPUs
	Bz	Epoch	Bz	Epoch	Bz	Epoch	
LLaVA	48	5	96	3	48	3	V100×8
Qwen2.5VL-3B	48	1	96	1	48	1	H200×4
Qwen2.5VL-7B	48	1	96	1	48	1	H200×4
Llama3.2-11B	48	1	96	1	48	1	V100×8
Gemma3-4B	48	1	96	1	48	1	H200×4
ChartInstruct	48	10	48	10	–	–	V100×8
Unichart	48	20	48	20	–	–	V100×8

former pathway of the base LLM. As a result, it may already contain information from outside the mask region M_t , propagated from earlier layers due to the inherent contextual accumulation across transformer layers.

Ideally, the inter-layer dependencies up to layer l should be explicitly accounted for to mitigate such information leakage, as this would allow us to trace how information from outside the mask region propagates across layers. In fact, Ferrando et al. proposed ALTI [2], which explicitly models token-to-token dependencies across layers and has been shown to improve attribution accuracy. However, such an approach requires additional token-to-token computations over feature dimensions, making it computationally expensive. Therefore, we do not incorporate this effect in our method in order to maintain computational efficiency and leave it as future work.

A2. Implementation details

Tab. 6 summarizes the hyperparameters used to train each VLM in our experiments. As pretrained models, we employed LLaVA¹, Qwen2.5VL-3B², Qwen2.5VL-7B³, Llama3.2-11B⁴, Gemma3-4B⁵, ChartInstruct⁶, and UniChart⁷ from the Hugging Face model hub. Due to limited computational resources, each model was trained on a different machine using V100 or H200 GPUs. However, for each model, both comparative methods and our proposed SFT were trained on identical hardware under the same conditions to ensure a fair comparison. Training was performed with DeepSpeed ZeRO3 [10], the Hugging Face Transformers library [12], and Pytorch [9]. We adopt the batch size and the number of training epochs for each model and task as shown in Tab. 6. Empirically, changing the global or per-

¹liuhaotian/llava-v1.5-7b

²Qwen/Qwen2.5-VL-3B-Instruct

³Qwen/Qwen2.5-VL-7B-Instruct

⁴meta-llama/Llama-3.2-11B-Vision-Instruct

⁵google/gemma-3-4b-it

⁶ahmed-masry/ChartInstruct-LLama2

⁷ahmed-masry/unichart-base-960

Table 7. Performance on other VLMs.

Model	Method	C2T		ChartQA		QA-CoT	
		Aug.	Hum.	Aug.	Hum.	Aug.	Hum.
LLaMA3.2-11B	<i>SFT</i>	96.4	87.6	93.0	55.4	93.2	75.4
	Ours	97.2	88.8	93.1	56.8	93.7	77.1
Gemma3-4B	<i>SFT</i>	95.0	85.6	90.1	55.3	89.8	72.3
	Ours	95.1	85.7	90.4	57.9	90.1	73.3

device batch size affects the performance of both the *SFT* baseline and our method; therefore, we fix the batch size and set the number of gradient accumulation steps to 1. For training all models, the initial learning rate was set to $5e-5$ and a linear decay schedule was applied. Additionally, we employed a warmup ratio of 3% of the total training steps. We first train the model using standard SFT and determine the optimal number of training epochs based on the validation score. For models whose performance degrades with additional training epochs, we set the number of epochs to one. Because our proposed SFT exhibits similar convergence behavior, we adopt the same training hyperparameters as those used in the baseline training. We freeze the vision encoder when training LLaVA and both Qwen models. When applying LLaVA in this work, we found that its image processor crops part of the chart during preprocessing. To avoid this, we resize chart images to a square resolution before feeding them into the image processor and the model. For Unichart and ChartInstruct, we set $\gamma = 1.0$. We use a larger weight K for the auxiliary loss $K \times \mathcal{L}_{\text{aux}}$ in QA-CoT. As shown in Tab. 1, the proportion of region-aligned tokens in QA-CoT is relatively low, making the auxiliary loss less effective when $K = 1.0$. Specifically, we set $K = 2.0$ for Qwen models and $K = 3.0$ for LLaVA, while using $K = 1.0$ for ChartQA across all models.

A3. Additional results

A3.1. Applicability to other VLMs

To demonstrate the applicability of SMAM to other VLMs, we additionally evaluated our method with Gemma3-4B [11] and Llama3.2-11B-Vision [4] as shown in Tab. 7. The result is consistent with the trends in the main paper and shows scalability to other models.

A3.2. Applicability to other domains

To demonstrate the applicability of SMAM beyond ChartQA, we further evaluate it on DocVQA [7] and InfographicVQA [8], using Acc. and Average Normalized Levenshtein Similarity (ANLS) metrics. Each model was trained for one epoch per dataset. The results are summarized in Tab. 8. For phrase-region alignment, we employ Paddle-OCR [1] instead of GPT to reduce the cost associated with API usage. Although this results in fewer align-

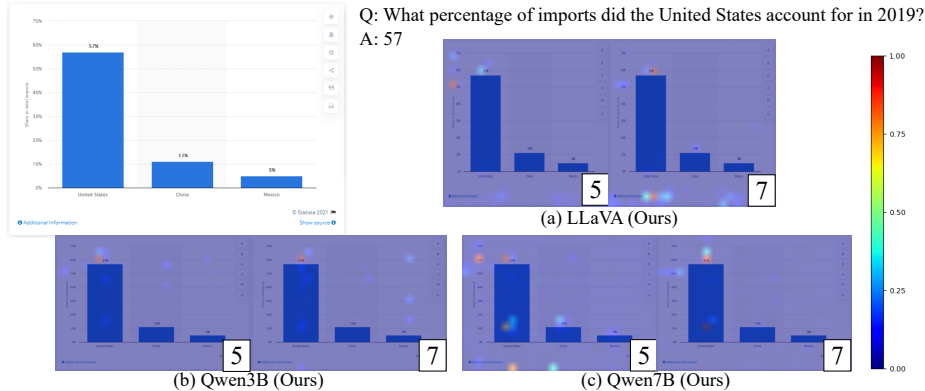


Figure 6. Heatmaps showing per-token vision-patch contributions, overlaid on the original charts. Panels (a), (b), and (c) show samples for LLaVA, Qwen3B, and Qwen7B, respectively, each fine-tuned with our method. The output token w_t is displayed below each heatmap.

Table 8. Results on other domain datasets.

Model	Method	DocVQA [7]		InfographicVQA [8]	
		Acc.	ANLS	Acc.	ANLS
LLaVA-7B	<i>SFT</i>	20.4	40.9	12.8	23.7
	Ours	20.9	40.8	14.8	27.0
Qwen2.5VL-3B	<i>SFT</i>	76.2	86.2	48.0	58.6
	Ours	77.8	87.5	47.6	58.2
Qwen2.5VL-7B	<i>SFT</i>	78.0	87.5	53.8	63.1
	Ours	79.0	88.9	55.2	64.0

Table 9. Evaluation of image captioning on Pew and Statista using the BLEU-4 metric.

Model	Method	Pew	Statista
LLaVA	<i>SFT</i>	13.2	26.0
	Ours	13.1	26.3
Qwen3B	<i>SFT</i>	13.5	30.8
	Ours	13.6	30.9
Qwen7B	<i>SFT</i>	15.6	32.2
	Ours	15.8	32.3

ments due to OCR limitations, SMAM still demonstrates overall effectiveness on these domains, except for a performance drop for Qwen3B on InfographicVQA.

A3.3. Evaluation on the chart captioning task

We also generated phrase-to-mask aligned data for the Pew and Statista captioning datasets [5]. Although these datasets were included in the main paper to evaluate the quality of our data-generation pipeline, captioning performance itself was not discussed. Here, we report the captioning results obtained using the same fine-tuning procedure as in the main experiments. Tab. 9 shows the BLEU-4 scores for LLaVA, Qwen3B, and Qwen7B. Across all models, our method does not show clear improvements over the

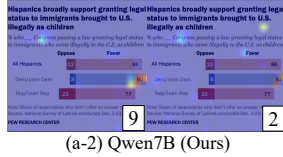
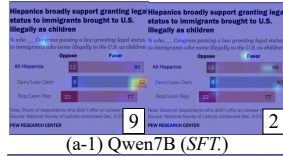
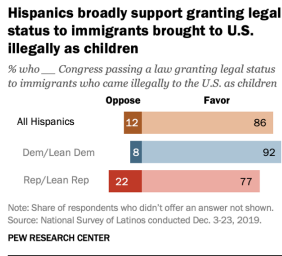
SFT baseline. BLEU-4 measures sentence-level similarity, whereas our method is designed to improve phrase-level grounding. As a result, improvements in phrase-level grounding, as demonstrated in the main paper, do not necessarily translate into gains on sentence-level metrics such as BLEU. This makes such improvements difficult to observe directly. In addition, the data statistics in Tab. 1 indicate that only about one-fifth to one-sixth of the tokens in the captioning datasets have phrase-to-mask alignments. This means that SMAM can apply its grounding objective to only a small portion of the caption tokens, which further limits the potential impact on BLEU. We present this result in the supplementary material for completeness.

A3.4. Visualization of vision-patch contributions

Visualization of per-patch contributions across models:

Fig. 6 visualizes vision-patch contributions as heatmaps for a bar-chart example across different models. Consistent with Fig. 4 in the main paper, all models show high contribution values around the region where “57” appears, which matches the generated phrase. For LLaVA, the heatmap for predicting the token “5” also shows strong activation near the “50” tick label on the y-axis. In the per-patch contribution computation based on LLaVA’s representations, visually similar digit regions tend to show non-negligible scores, resulting in false-positive activations on other occurrences of the digit “5”. For Qwen3B, only the region corresponding to “57” exhibits high contribution, indicating more precise localization under the same computation procedure. For Qwen7B, the computed contributions concentrate primarily on the correct “57” region, while also showing moderate activation near the “60” axis label and the corresponding bar. This suggests that the model leverages nearby axis markings, such as “60”, as contextual information to support the prediction of “57”, while still distinguishing it from visually similar digit regions. Overall, these observations show that different models exhibit

Q: What is the value of Favor on the blue bar?
A: 92



Q: What is the highest webcam penetration rate?
A: 48

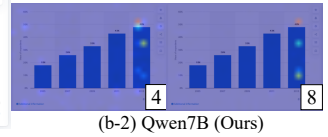
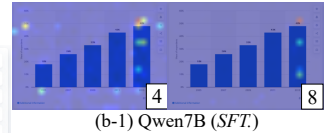
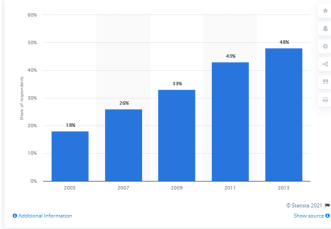
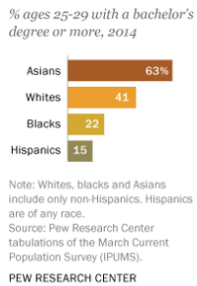
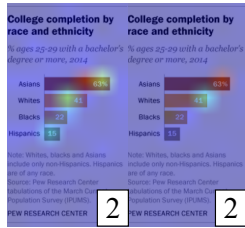


Figure 7. Heatmaps of per-token vision-patch contributions for two samples generated by Qwen7B, overlaid on the original charts. Panels (a-1) and (a-2) compare the *SFT* baseline with our method for the first sample. Panels (b-1) and (b-2) show the corresponding comparison for the second sample. The output token w_t is displayed below each heatmap.

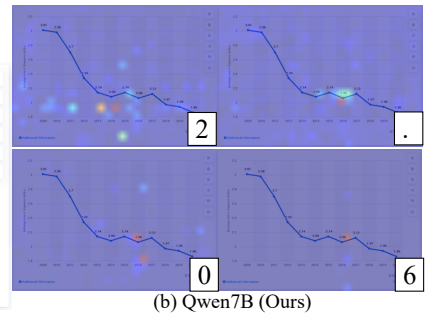
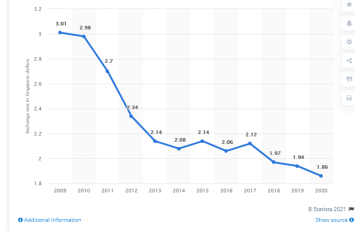
College completion by race and ethnicity



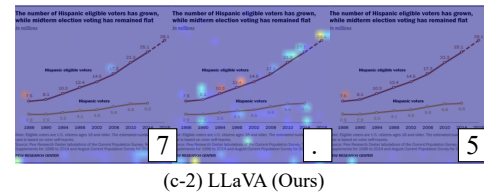
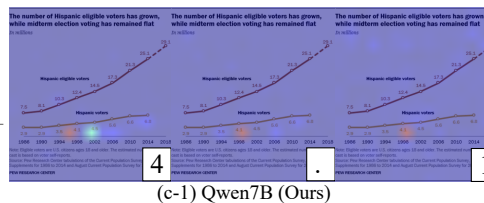
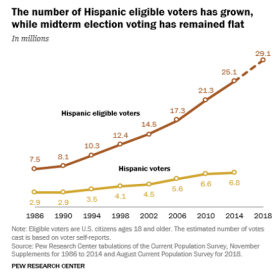
Q: What is the difference between Asians and Whites degree distribution?
A: 22



Q: What is the data value in 2016?
A: 2.06



Q: Find missing value- 2.9, 2.9, 3.5, 4.5, 5.6, 6.6, 6.8?
A: 4.1
Qwen7B (ours): 4.1
LLaVA (ours): 7.5



Q: What was the 4th most popular emotion?
A: Inspired

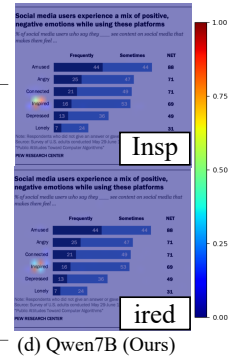
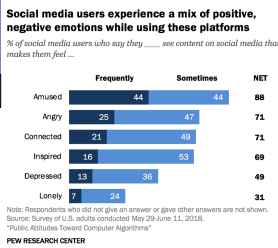


Figure 8. Additional heatmap visualizations. (a), (b), (c-1), and (d) show results from Qwen7B, and (c-2) shows an incorrect example from LLaVA.

distinct patterns of vision-patch contributions during token prediction. For LLaVA, the heatmaps reveal a tendency toward surface-level matches between phrases and visually similar regions, whereas Qwen7B exhibits more semantically grounded behavior, combining evidence from the correct region with contextually informative cues.

Visualization comparing *SFT* and our method: Fig. 7 shows the difference in heatmap behavior between the *SFT* baseline and our method for Qwen7B. Consistent with the observations in Fig. 4 of the main paper, the contribution pattern becomes sharper around the correct region. How-

ever, the heatmaps do not collapse into an overly narrow focus. Rather, as the model adapts to the ChartQA training dataset, it continues to contain contributions in regions needed for the answer generation while placing relatively stronger emphasis on the correct region. This behavior aligns with the design of our SMAM objective, which encourages the model to base its prediction on the essential evidence within the phrase-aligned region while still leveraging useful contextual cues from the rest of the image. The resulting heatmaps therefore exhibit the same qualitative pattern observed in the main paper (Fig. 4), where the cor-

rect region is emphasized without losing relevant surrounding context.

Additional visualization examples: Finally, we present several additional visualization examples. Fig. 8 (a) shows a Qwen7B case that requires subtraction. When generating the answer “22”, the heatmap does not place high contribution scores on the superficial “22” region in the chart. Instead, it highlights the regions corresponding to “63” and “41”, which provide the actual arithmetic evidence. Fig. 8 (b), (c-1), and (c-2) show samples involving line plots. Because most examples in the main paper focused on bar charts, we include these cases to demonstrate that the visualization behaves similarly for plot-style figures. Fig. 8 (c-1) and (c-2) correspond to a missing-value problem that is somewhat challenging. Qwen7B answers correctly, whereas LLaVA fails. In the LLaVA case (c-2), the contribution map becomes scattered when predicting the token “.”, and unrelated locations receive non-negligible scores, indicating false-positive behavior under this contribution computation. Fig. 8 (d) illustrates a sample that requires predicting a word rather than a number. In the earlier numeric-reading samples, Qwen7B often made use of surrounding visual context for the predictions. In contrast, when the corresponding text region is visually unique, the heatmap becomes sharply concentrated on the area where “Inspired” appears.

In principle, vision-patch contribution maps can also be computed for C2T and QA-CoT, where the output sequences are longer. In practice, however, computing these heatmaps requires storing contributions for every output token, and our extended implementation based on the method of Ferrando et al. [3] became infeasible for long sequences due to memory limitations. For this reason, all heatmap evaluations and visualizations in this study are reported for the ChartQA task, whose short phrase-level outputs keep the computational cost manageable.

A4. Dataset details and examples

A4.1. GPT-based data generation: prompts and examples

We used the GPT API with the *o4-mini-2025-04-16* model to generate our dataset. The prompt for QA-CoT generation is shown in Fig. 9, and the prompts for Pew and Statista are shown in Fig. 10. For each sample, we submitted the alpha-blended image V_α together with its corresponding prompt, in which the actual text T is embedded, yielding the phrase-to-mask aligned text output T' . Examples of the inputs V_α and their generated responses are provided in Figs. 12, 13, and 14, for QA-CoT, Pew, and Statista samples, respectively. Since QA-CoT contains the same answer portion as ChartQA, we reuse the answer part of QA-CoT for the ChartQA task.

All GPT requests were made via the batch API to minimize cost: approximately USD 60 for Pew, USD 160 for Statista, and USD 160 for QA-CoT generation. To evaluate the per-patch contribution heatmaps shown in Tab. 4 of the main paper, the ground-truth masks for the answer phrases are required. Thus we also augmented the ChartQA test split using the prompt in Fig. 11, at an additional cost of approximately USD 17.

A4.2. Dataset used for C2T training

Finally, Fig. 15 presents samples of the dataset used to train the C2T task. We utilized the bounding-box annotations and corresponding numeric label annotations provided by ChartQA [6], aligning each annotation with its corresponding table value for training. We did not use GPT API in this setting. Although these annotations are supplied in the bounding-box format, we convert them into a mask format. As illustrated in the figure, the original dataset’s annotation protocol (shown in Fig. 15 (a-1) and (b-1)) differs from ours (shown in (a-3) and (b-3)): we generate masks over text regions, such as data labels, using Hi-SAM, whereas the original annotations target chart elements such as bars and line plots, and occasionally some annotations are missing as shown in Fig. 15 (a-1). Nevertheless, as we demonstrated in our paper, our method performs effectively across both annotation formats. Moreover, because our approach is applied only to the output tokens w_t that have a valid mask-to-phrase correspondence, it remains robust even when the provided annotations are partial or incomplete.

A5. Limitation

In this work, we mainly focus on phrase–region alignment on textual regions when constructing the dataset, since such alignment can be quantitatively evaluated using OCR or GPT-based verification. However, in document and diagram understanding, it is often desirable to align phrases with non-textual elements beyond literally corresponding text regions. Currently, very few datasets provide phrase–region alignment for non-textual elements in a VQA-style format. It remains difficult to systematically evaluate the quality of such alignments, even when new datasets are introduced. Although we partially demonstrate that our method can utilize annotations for bars or plot instances provided in ChartQA through the C2T task, verifying whether the method generalizes reliably to broader non-textual elements remains future work.

In addition, for a more faithful SMAM computation, token-to-token dependencies across layers need to be considered, as discussed in Sec. A1. We leave this for future investigation.

References

- [1] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiakuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 2
- [2] Javier Ferrando, Gerard I Gállego, and Marta R Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proc. of EMNLP*, pages 8698–8714, 2022. 2
- [3] Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. Explaining how transformers use context to build predictions. In *Proc. of ACL*, pages 5486–5513, 2023. 1, 5
- [4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2
- [5] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. 3
- [6] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 5
- [7] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proc. of WACV*, pages 2200–2209, 2021. 2, 3
- [8] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proc. of WACV*, pages 1697–1706, 2022. 2, 3
- [9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. *NeurIPS Workshop on Automatic Differentiation*, 2017. 2
- [10] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 2
- [11] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2
- [12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 2

I've provided a chart image with numbered regions, a question, and a reasoning sentence, and I want you to modify the sentences to derive the conclusion. The chart contains several colored regions, each labeled with a numbered marker ([1], [2], etc.). These markers correspond to areas of text in the chart. I want you to align each numbered region with the matching phrase

```
in the given sentences and verify whether the text in
the region matches that phrase.

# The given sentences:
The question for the chart is:
{question}
The reasoning for deriving the conclusion is:
{reasoning}
The conclusion is:
{ans}

# Requirements for the rewriting sentences
- Tag each phrase (word or value) in the given
sentences with the corresponding marker using <MARK_XX
> and </MARK_XX> tags (for example, <MARK_7>phrase</
MARK_7> for marker [7]).
- Do not tag any phrases that do not correspond to a
marker. If markers overlap or are hard to distinguish,
omit tagging for clarity.
- You do not need to search the web for this task.
Focus only on the provided chart.
- For this task, first examine the text in each
numbered region of the chart and then align it with
the corresponding phrase.

# Requirements for the response
- Wrap all rewritten sentences using <ans> and </ans>
tags.
- In the response, rewrite the reasoning and
conclusion, wrapping each part with <REASONING>...</
REASONING> and <CONCLUSION>...</CONCLUSION>.
- Do not change the original content; just add <
MARK_XX> and </MARK_XX> tags.
- Don't include any other sentences between the <ans>
and </ans> tags.

# The response format is as follows:
<ans>
<REASONING>rewritten reasoning here</REASONING>
<CONCLUSION>rewritten conclusion here</CONCLUSION>
</ans>
```

Figure 9. Prompt used for QA-CoT dataset. {question}, {reasoning}, {ans} are replaced by the actual sentences in the text T for each sample.

```
I've provided a chart image and its caption. The chart
contains several colored regions, each labeled with a
numbered marker ([1], [2], etc.). These markers
correspond to areas of text in the chart. I want to
align each numbered region with the matching phrase in
the caption and verify whether the text in the region
matches that phrase.

# The given caption:
{caption}

# Requirements for the rewriting the caption
- Tag each phrase (word or value) in the caption with
the corresponding marker using <MARK_XX> and </MARK_XX
> tags (for example, <MARK_7>phrase</MARK_7> for
marker [7]).
- Do not tag any phrases that do not correspond to a
marker. If markers overlap or are hard to distinguish,
omit tagging for clarity.
- You do not need to search the web for this task.
Focus only on the provided chart.
- For this task, first examine the text in each
numbered region of the chart and then align it with
the corresponding phrase.

# Requirements for the response
- you must wrap the rewritten caption using <ans> and
</ans> tag.
- Don't change the original caption; just add <MARK_XX
```

```

> and </MARK_XX> tags to the caption.
- Don't include any other sentences between the <ans>
and </ans> tags.

# Response format
<ans>[rewritten caption here.]</ans>

```

Figure 10. Prompt used for Pew and Statista dataset. {caption} is replaced by the actual caption in the text T for each sample.

```

I've provided a chart image and its QA pair. The chart
contains several colored regions, each labeled with a
numbered marker ([1], [2], etc.). These markers
correspond to areas of text in the chart. I want you
to align each numbered region with the matching phrase
in the given sentences and verify whether the text in
the region matches that phrase.

# The given question:
{question}
# The given answer:
{ans}

# Requirements for the rewriting the answer
- Tag each phrase (word or value) in the given
sentences with the corresponding marker using <MARK_XX>
> and </MARK_XX> tags (for example, <MARK_7>phrase</
MARK_7> for marker [7]).
- Do not tag any phrases that do not correspond to a
marker. If markers overlap or are hard to distinguish,
omit tagging for clarity.
- You do not need to search the web for this task.
Focus only on the provided chart.
- For this task, first examine the text in each
numbered region of the chart and then align it with
the corresponding phrase.

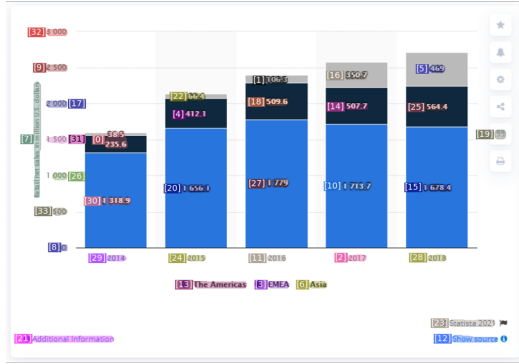
# Requirements for the response
- you must wrap the rewritten answer using <ans> and
</ans> tag.
- Don't change the original answer; just add <MARK_XX>
and </MARK_XX> tags to the answer.
- Don't include any other sentences between the <ans>
and </ans> tags.
- If the answer does not match any markers, you don't
need to add tags.

# Response format
<ans>[rewritten answer here.]</ans>

```

Figure 11. Prompt used for ChartQA test data. {question} and {ans} are replaced by the actual sentences in the text T for each sample.

Original question answer pair T from the dataset.



Alpha-blended image: V_α

Question:

What was the retail net sales generated by Michael Kors' EMEA region in dollars in 2018?

Answer:

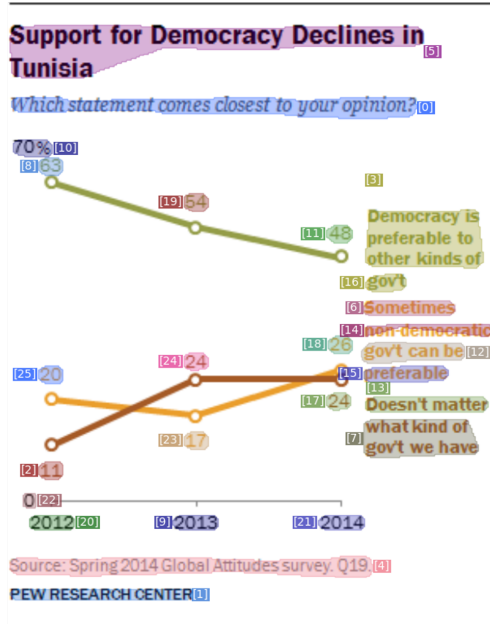
<REASONING> To find the retail net sales for the EMEA region in 2018, I will look at the 2018 bar and identify the segment representing EMEA, which is colored in dark blue. The labeled value within this segment will give the sales figure specifically for the EMEA region in that year.

</REASONING>\n<CONCLUSION> 564.4 </CONCLUSION>

Parse-to-mask aligned T' obtained as the GPT's response

<REASONING>To find the retail net sales for the <MARK_3>EMEA</MARK_3> region in <MARK_28>2018</MARK_28>, I will look at the <MARK_28>2018</MARK_28> bar and identify the segment representing <MARK_3>EMEA</MARK_3>, which is colored in dark blue. The labeled value within this segment will give the sales figure specifically for the <MARK_3>EMEA</MARK_3> region in that year.</REASONING>\n<CONCLUSION><MARK_25>564.4</MARK_25></CONCLUSION>\n

Original question answer pair T from the dataset.



Alpha-blended image: V_α

Question:

When does the percentage of people who prefer democracy reach the peak?

Answer:

<REASONING> To identify when support for democracy is at its peak, I will focus on the line representing "Democracy is preferable to other kinds of gov't." Observing this line, it begins at 63% in 2012, which is the highest percentage point across the displayed years. The trend indicates a decline thereafter, with percentages falling to 54% in 2013 and 48% in 2014. Therefore, the peak is clearly at the starting point of the timeline, in 2012.

</REASONING>\n<CONCLUSION> 2012 </CONCLUSION>

Parse-to-mask aligned T' obtained as the GPT's response

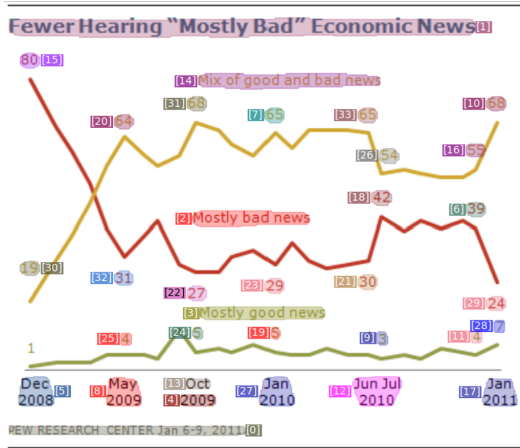
<REASONING>To identify when support for democracy is at its peak, I will focus on the line representing "<MARK_3>Democracy is preferable to other kinds of gov't." Observing this line, it begins at <MARK_8>63%</MARK_8> in <MARK_20>2012</MARK_20>, which is the highest percentage point across the displayed years. The trend indicates a decline thereafter, with percentages falling to <MARK_19>54%</MARK_19> in <MARK_9>2013</MARK_9> and <MARK_11>48%</MARK_11> in <MARK_21>2014</MARK_21>. Therefore, the peak is clearly at the starting point of the timeline, in <MARK_20>2012</MARK_20>.</REASONING>\n<CONCLUSION><MARK_20>2012</MARK_20></CONCLUSION>\n

Figure 12. Samples of GPT inputs consisting of the text T and the alpha-blended image V_α , along with the corresponding GPT-generated outputs T' for the QA-CoT dataset.

Original question answer pair T from the dataset.

Question:
Write a summary for this chart.

Answer:
Public perceptions of the tone of economic news have become much less negative over the past month. In fact, the percentage saying they are hearing mostly bad news about the economy has dropped to its lowest point since the question was first asked in December 2008. Currently, 24% say they are hearing mostly bad news, down 15 points from 39% in early December. The proportion saying they are hearing a mix of good and bad news has jumped from 55% last month to 68% in the new survey. However, as has been the case for the past two years, very few say they are hearing mostly good economic news; just 7% say this, which is little changed from 4% in December.



Alpha-blended image: V_α

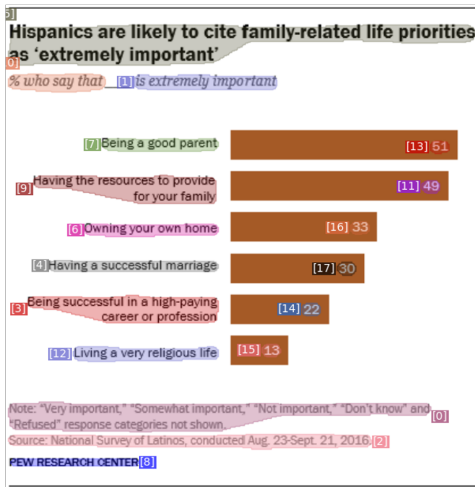
Phrase-to-mask aligned T' obtained as the GPT's response.

Public perceptions of the tone of economic news have become much less negative over the past month. In fact, the percentage saying they are hearing mostly bad news about the economy has dropped to its lowest point since the question was first asked in December 2008. Currently, 24% say they are hearing mostly bad news, down 15 points from 39% in early December. The proportion saying they are hearing a mix of good and bad news has jumped from 55% last month to 68% in the new survey. However, as has been the case for the past two years, very few say they are hearing mostly good economic news; just 7% say this, which is little changed from 4% in December.

Original question answer pair T from the dataset.

Question:
Write a summary for this chart.

Answer:
The life goals of Latinos overlap with some of these ingredients of the American dream. The two goals rated most highly by Latinos were being a good parent (51%) and having the resources to provide for their family (49%). Owning a home – often seen as a key part of the American dream – was cited by 33% of Hispanics as an extremely important life goal for them. Other top goals for Hispanics included having a successful marriage (30%) and being successful in a high-paying career or profession (22%).



Alpha-blended image: V_α

Phrase-to-mask aligned T' obtained as the GPT's response.

The life goals of Latinos overlap with some of these ingredients of the American dream. The two goals rated most highly by Latinos were being a good parent (51%) and having the resources to provide for their family (49%). Owning a home – often seen as a key part of the American dream – was cited by 33% of Hispanics as an extremely important life goal for them. Other top goals for Hispanics included having a successful marriage (30%) and being successful in a high-paying career or profession (22%).

Figure 13. Samples of GPT inputs consisting of the text T and the alpha-blended image V_α , along with the corresponding GPT-generated outputs T' for Pew dataset.

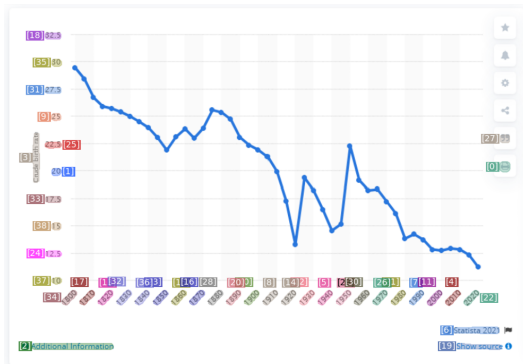
Original question answer pair T from the dataset.

Question:

Write a summary for this chart.

Answer:

In France , the crude birth rate in 1800 was 29.4 live births per thousand people , meaning that 2.9 percent of the population had been born in that year . In the first half of the nineteenth century France 's crude birth rate dropped from it 's highest recorded level of 29.4 in 1800 , to 21.9 by 1850 . In the second half of the 1800s the crude birth rate rose again , to 25.5 in 1875 , as the Second Republic and Second Empire were established , which was a time of economic prosperity and the modernization of the country . From then until 1910 there was a gradual decline , until the First World War caused a huge decline , resulting in a record low crude birth rate of 13.3 by 1920 (the figures for individual years fell even lower than this) . The figure then bounced back in the early 1920s , before then falling again until the Second World War . After the war , France experienced a baby boom , where the crude birth rate reached 22.2 , before it dropped again until the 1980s , and since then it has declined slowly . The crude birth rate of France is expected to reach a new , record low of 11.2 in 2020 .



Alpha-blended image: V_α

Parse-to-mask aligned T' obtained as the GPT's response

In France, the crude birth rate in <MARK_34>1800</MARK_34> was 29.4 live births per thousand people, meaning that 2.9 percent of the population had been born in that year. In the first half of the nineteenth century France's crude birth rate dropped from its highest recorded level of 29.4 in <MARK_34>1800</MARK_34>, to 21.9 by <MARK_16>1850</MARK_16>. In the second half of the 1800s the crude birth rate rose again, to 25.5 in 1875, as the Second Republic and Second Empire were established, which was a time of economic prosperity and the modernization of the country. From then until 1910 there was a gradual decline, until the First World War caused a huge decline, resulting in a record low crude birth rate of 13.3 by <MARK_26>1920</MARK_26> (the figures for individual years fell even lower than this). The figure then bounced back in the early 1920s, before then falling again until the Second World War. After the war, France experienced a baby boom, where the crude birth rate reached 22.2, before it dropped again until the 1980s, and since then it has declined slowly. The crude birth rate of France is expected to reach a new, record low of 11.2 in <MARK_22>2020</MARK_22>.

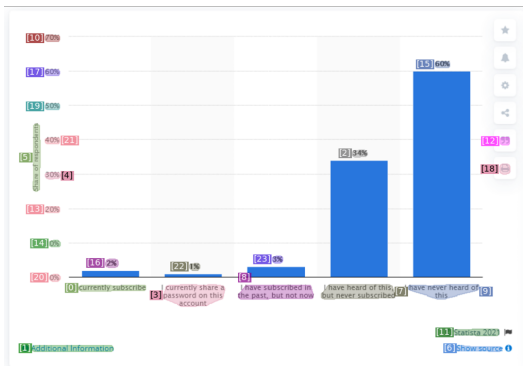
Original question answer pair T from the dataset.

Question:

Write a summary for this chart.

Answer:

Awareness of TV streaming site Philo with live TV is low among adults in the United States , with 60 percent of respondents having never heard of the platform as of May 2020 . Although 34 percent of those surveyed had heard of the service but never subscribed , just two percent currently subscribed to Philo .



Alpha-blended image: V_α

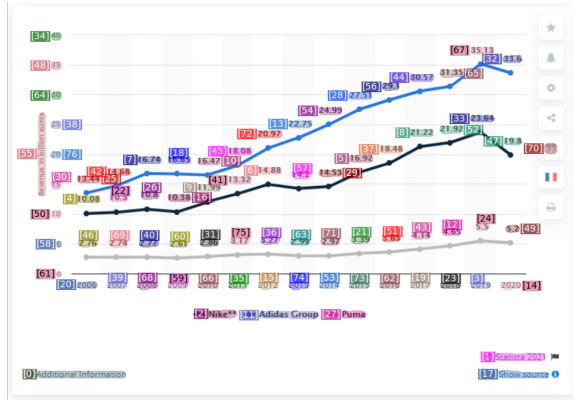
Parse-to-mask aligned T' obtained as the GPT's response

Awareness of TV streaming site Philo with live TV is low among adults in the United States , with <MARK_15>60 percent</MARK_15> of respondents having <MARK_9>never heard of the platform</MARK_9> as of May 2020 . Although <MARK_2>34 percent</MARK_2> of those surveyed had <MARK_7>heard of the service but never subscribed</MARK_7>, just <MARK_16>two percent</MARK_16> <MARK_0>currently subscribed to Philo</MARK_0>.

Figure 14. Samples of GPT inputs consisting of the text T and the alpha-blended image V_α , along with the corresponding GPT-generated outputs T' for Statista dataset.



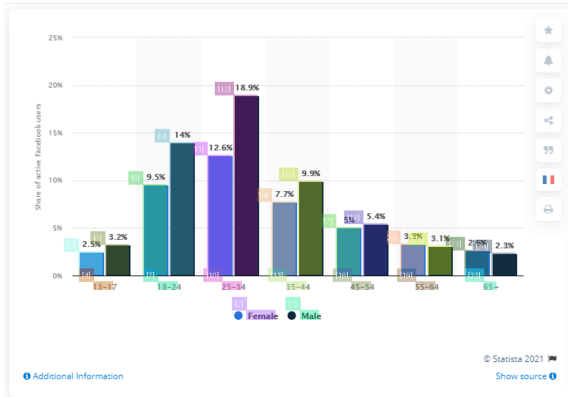
(a-1) Annotations from the original dataset.



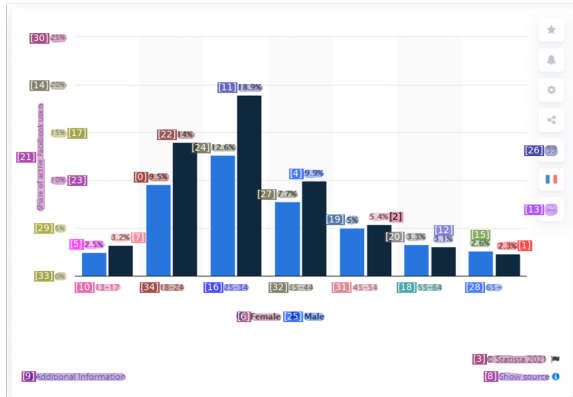
(a-3) Our mask annotations.

Characteristic | <MARK_2>Nike*</MARK_2> | <MARK_3>Adidas Group</MARK_3> | <MARK_4>Puma</MARK_4>
 <MARK_5>2020</MARK_5> | 33.6 | 19.8 | 5.2
 <MARK_9>2019</MARK_9> | 35.13 | 23.64 | 5.5
 <MARK_13>2018</MARK_13> | 31.35 | 21.92 | 4.65
 <MARK_17>2017</MARK_17> | 30.57 | 21.22 | 4.14
 <MARK_21>2016</MARK_21> | 29.1 | 18.48 | 3.63
 <MARK_25>2015</MARK_25> | 27.51 | 16.92 | 3.39
 <MARK_29>2014</MARK_29> | 24.99 | 14.53 | 2.97
 <MARK_33>2013</MARK_33> | 22.75 | 14.2 | 2.99
 <MARK_37>2012</MARK_37> | <MARK_38>20.97</MARK_38> | <MARK_39>14.88</MARK_39> | <MARK_40>3.27</MARK_40>
 <MARK_41>2011</MARK_41> | <MARK_42>18.08</MARK_42> | <MARK_43>13.32</MARK_43> | <MARK_44>3.17</MARK_44>
 <MARK_45>2010</MARK_45> | <MARK_46>16.47</MARK_46> | <MARK_47>11.99</MARK_47> | <MARK_48>2.86</MARK_48>
 <MARK_49>2009</MARK_49> | <MARK_50>16.65</MARK_50> | <MARK_51>10.38</MARK_51> | <MARK_52>2.61</MARK_52>
 <MARK_53>2008</MARK_53> | 16.74 | <MARK_55>10.8</MARK_55> | <MARK_56>2.77</MARK_56>
 <MARK_57>2007</MARK_57> | <MARK_58>14.68</MARK_58> | <MARK_59>10.3</MARK_59> | 2.74
 <MARK_61>2006</MARK_61> | <MARK_62>13.44</MARK_62> | <MARK_63>10.08</MARK_63> | 2.76

(a-2) Text data on the table extraction aligned to the original annotation.



(b-1) Annotations from the original dataset.



(b-3) Our mask annotations.

Characteristic | <MARK_2>Female</MARK_2> | <MARK_3>Male</MARK_3>
 <MARK_4>13-17</MARK_4> | <MARK_5>2.5%</MARK_5> | <MARK_6>3.2%</MARK_6>
 <MARK_7>18-24</MARK_7> | <MARK_8>9.5%</MARK_8> | <MARK_9>14%</MARK_9>
 <MARK_10>25-34</MARK_10> | <MARK_11>12.6%</MARK_11> | <MARK_12>18.9%</MARK_12>
 <MARK_13>35-44</MARK_13> | <MARK_14>7.7%</MARK_14> | <MARK_15>9.9%</MARK_15>
 <MARK_16>45-54</MARK_16> | <MARK_17>5%</MARK_17> | <MARK_18>5.4%</MARK_18>
 <MARK_19>55-64</MARK_19> | <MARK_20>3.3%</MARK_20> | <MARK_21>3.1%</MARK_21>
 <MARK_22>65+</MARK_22> | <MARK_23>2.6%</MARK_23> | <MARK_24>2.3%</MARK_24>

(b-2) Text data on the table extraction aligned to the original annotation.

Figure 15. The samples from the dataset used for C2T task. (a-1) shows the original ChartQA annotation; (a-2) displays the source table with each cell value aligned to the annotations in (a-1); and (a-3) presents our mask annotation showing different annotation protocol from (a-1). The same applies to (b).