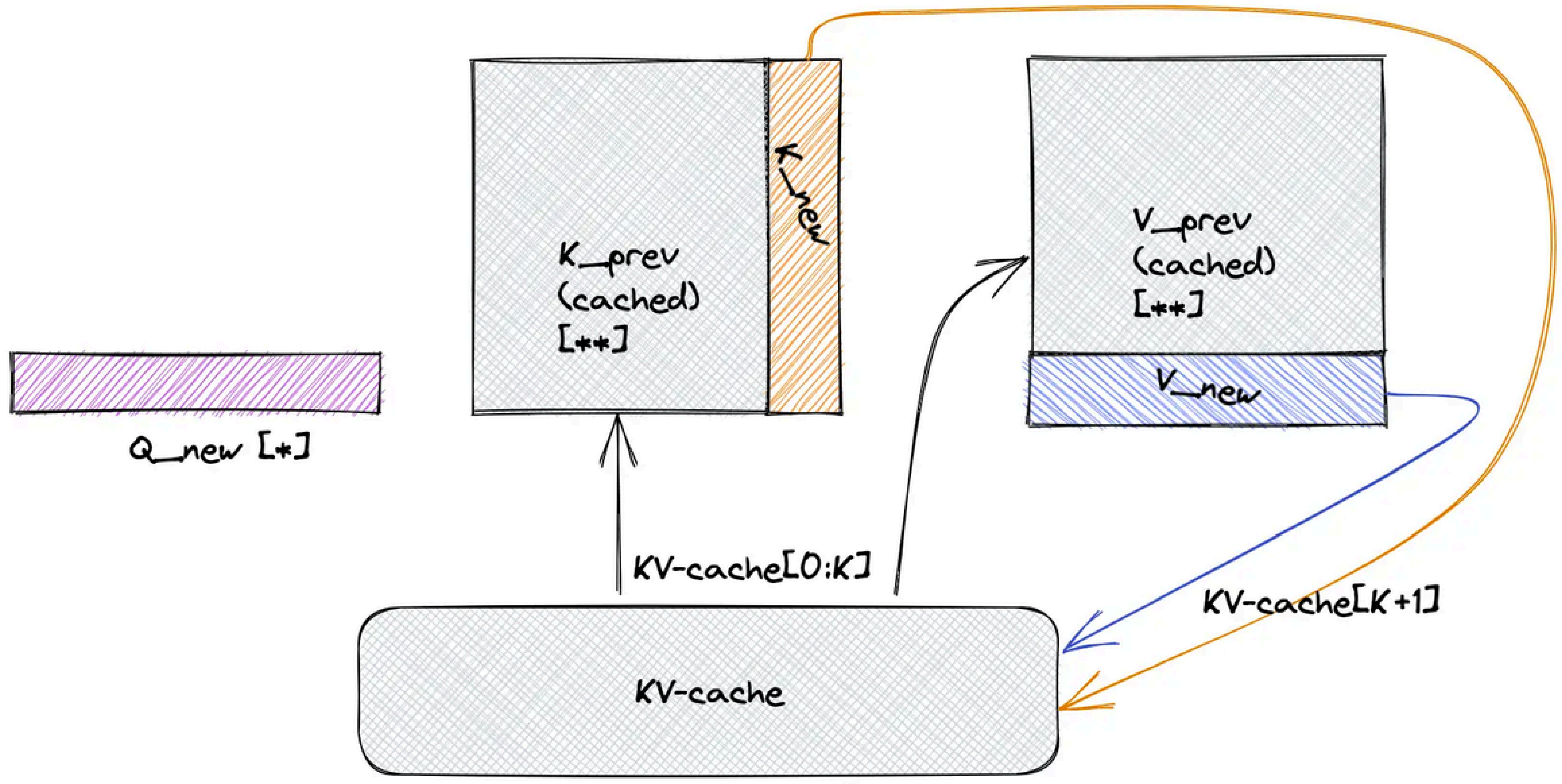
Technical and Architectural Enhancements in Gemma 3



Notes:

----

* When processing token[K], we only need the K'th row of Q

** When processing token[K], we require the full K & V tensors, but we can mostly reuse the cached values (This enables skipping the computation of K & V

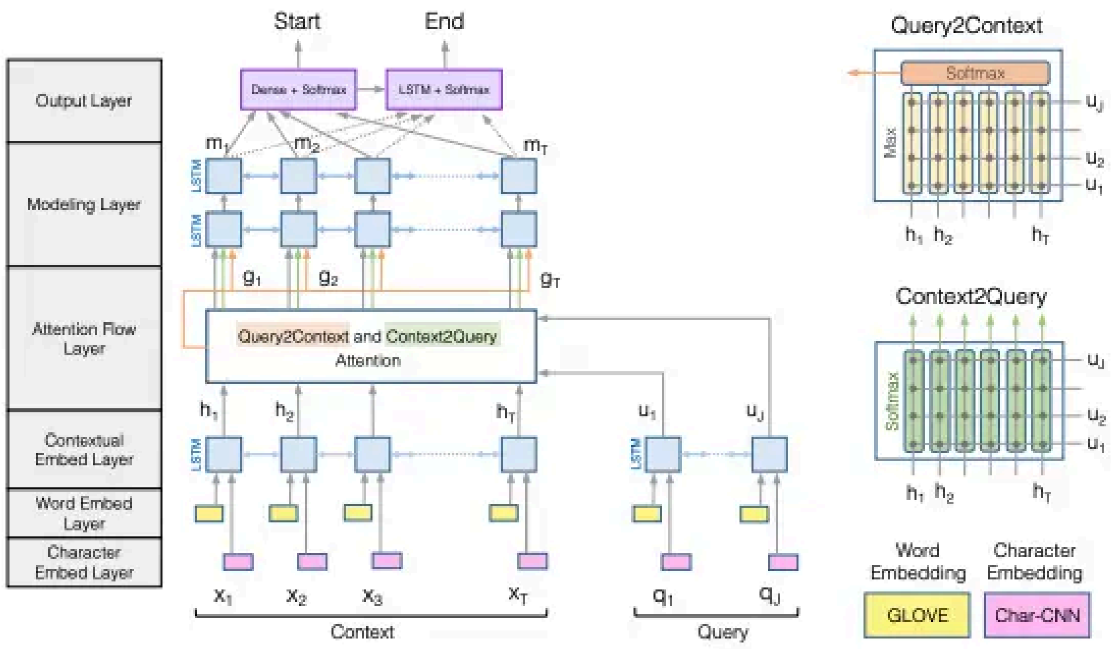
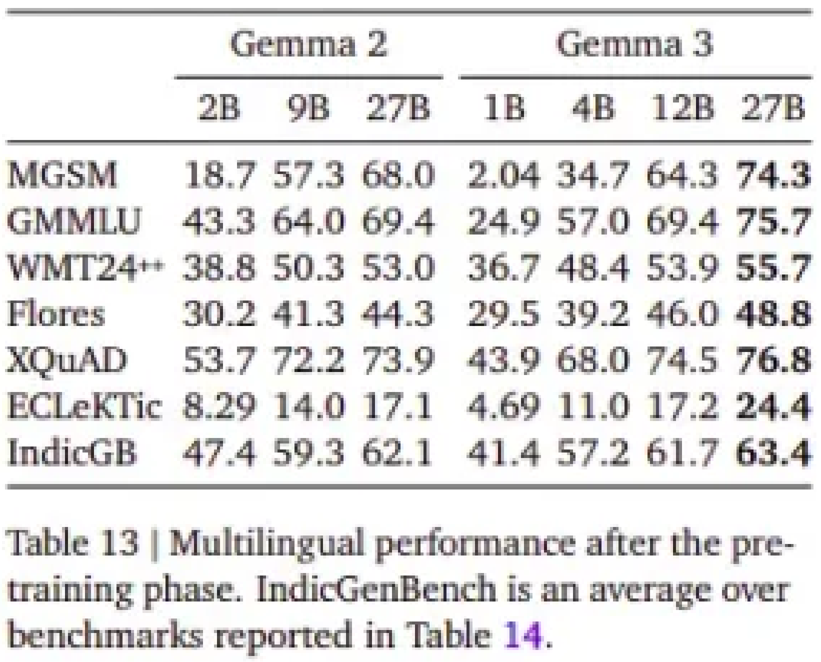


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

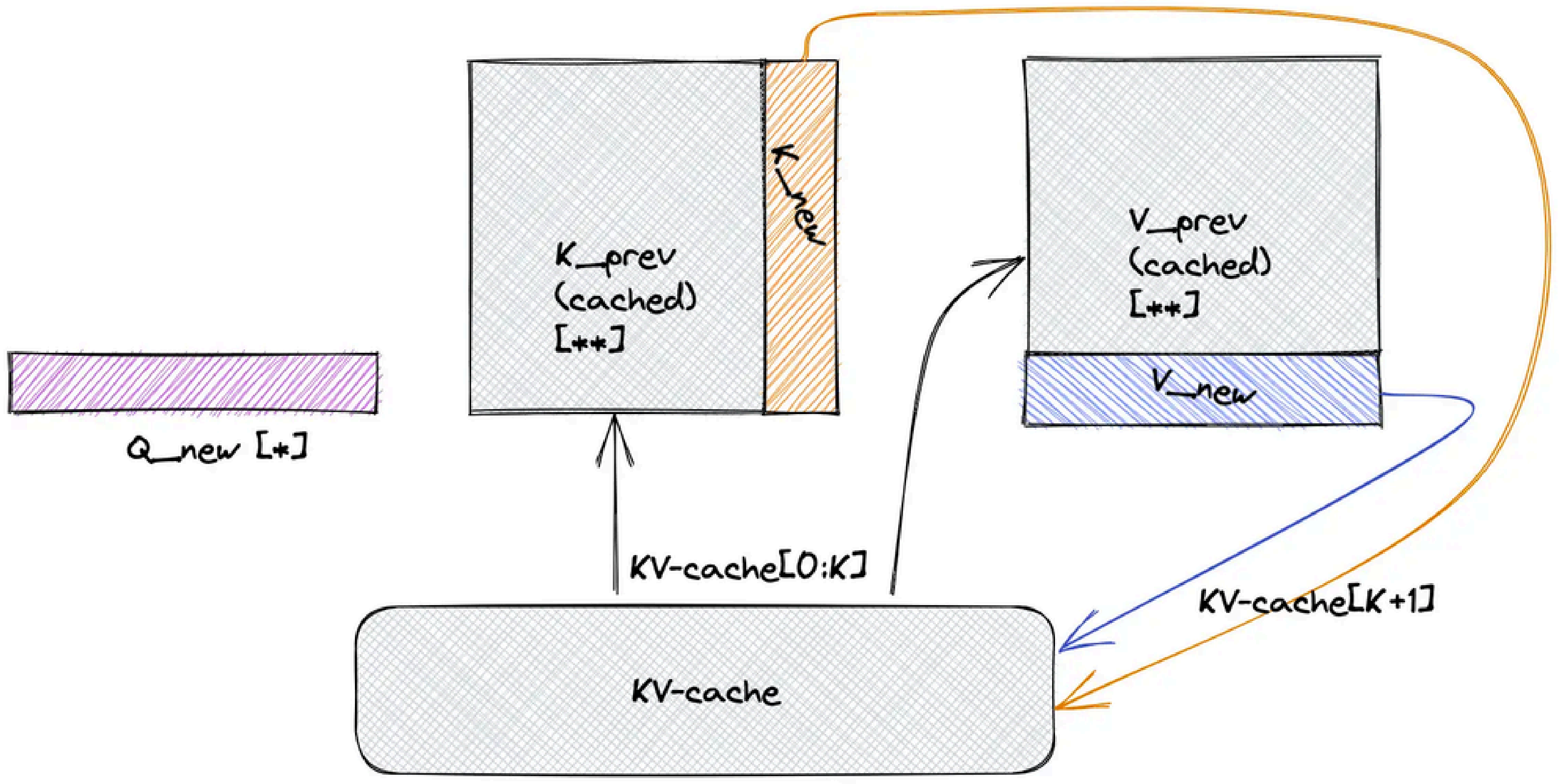


	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
MGSM	18.7	57.3	68.0	2.04	34.7	64.3	**74.3**
GMMLU	43.3	64.0	69.4	24.9	57.0	69.4	**75.7**
WMT24++	38.8	50.3	53.0	36.7	48.4	53.9	**55.7**
Flores	30.2	41.3	44.3	29.5	39.2	46.0	**48.8**
XQuAD	53.7	72.2	73.9	43.9	68.0	74.5	**76.8**
ECLeKTic	8.29	14.0	17.1	4.69	11.0	17.2	**24.4**
IndicGB	47.4	59.3	62.1	41.4	57.2	61.7	**63.4**

Table 13 | Multilingual performance after the pre-training phase. IndicGenBench is an average over benchmarks reported in Table 14.

Longer Context Length

- **Scaling Without Re-training from Scratch**: Models are initially pre-trained with 32K sequences. For the 4B, 12B, and 27B variants, the context length is efficiently scaled to 128K tokens post pre-training, saving significant compute.
- **Enhanced Positional Embeddings**: The RoPE (Rotary Positional Embedding) base frequency is upgraded from 10K in Gemma 2 to 1 M in Gemma 3 and then scaled by a factor of 8. This enables the models to maintain high performance even with extended context.
- **Optimized KV Cache Management**: By interleaving multiple local attention layers (with a sliding window of 1024 tokens) between global layers (at a 5:1 ratio), Gemma 3 dramatically reduces the KV cache memory overhead during inference from around 60% in global-only setups to less than 15%.



Notes:
----
* When processing token[K], we only need the K'th row of Q
** When processing token[K], we require the full K & V tensors, but we can mostly reuse the cached values
(This enables skipping the computation of K & V

Multimodality

- **Vision Encoder Integration**: Gemma 3 leverages the SigLIP image encoder to process images. All images are resized to a fixed 896×896 resolution for consistency. To handle non-square aspect ratios and high-resolution inputs, an adaptive "pan and scan" algorithm crops and resizes images on the fly, ensuring that critical visual details are preserved.

- **Distinct Attention Mechanisms**: While text tokens use one-way (causal) attention, image tokens receive bidirectional attention. This allows the model to build a complete and unrestricted understanding of visual inputs while maintaining efficient text processing.

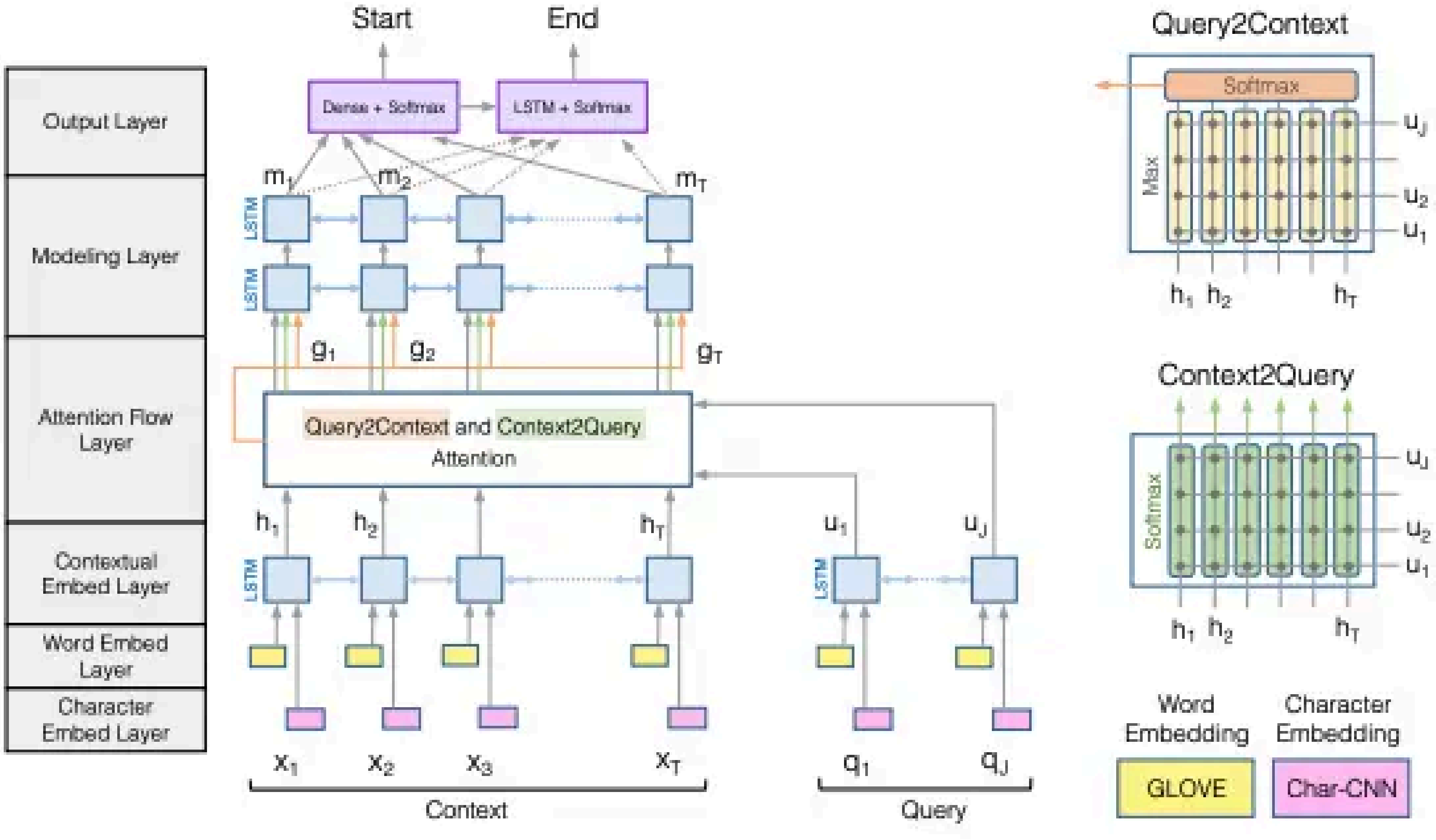
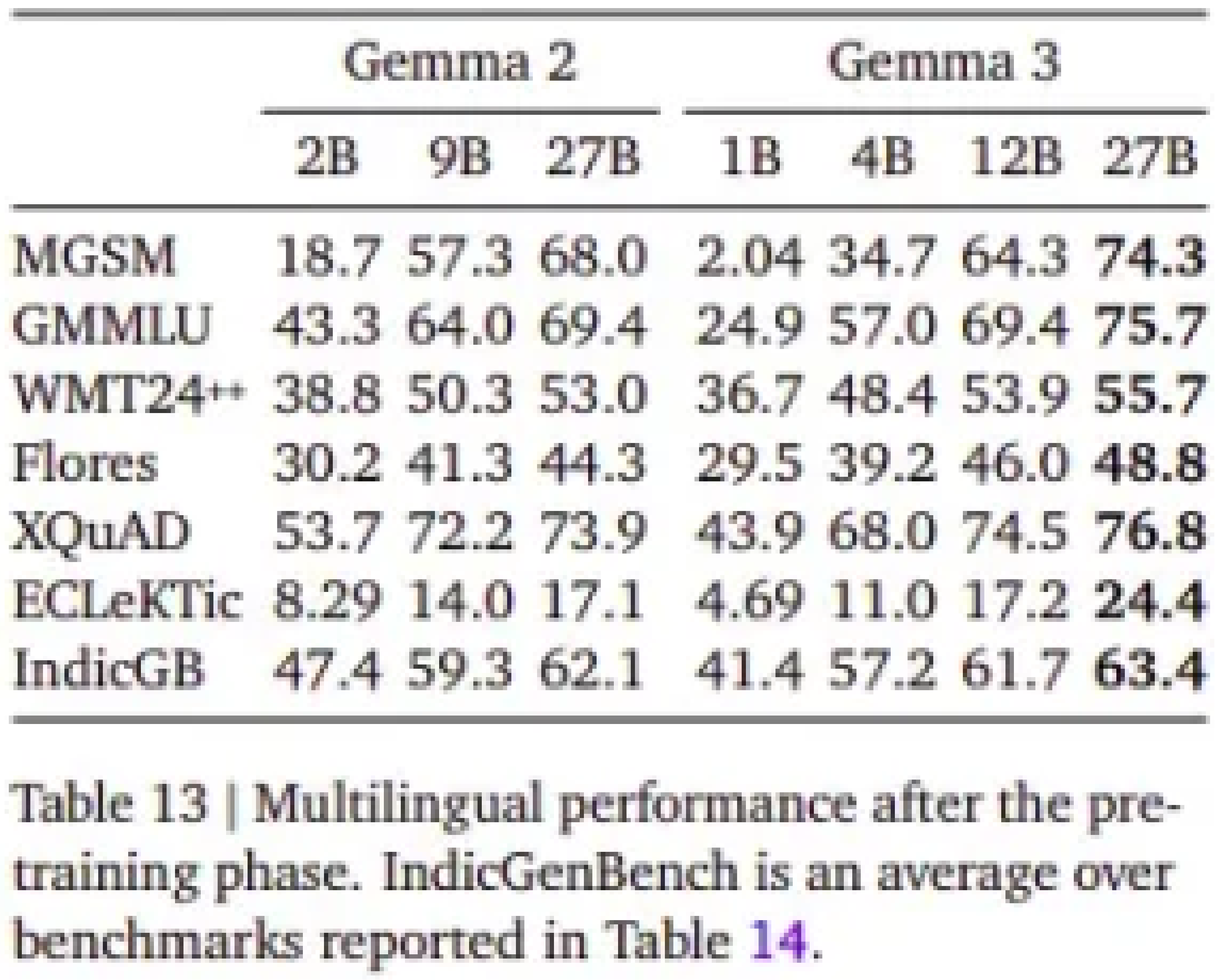


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Multilinguality

- Expanded Data and Tokenizer Improvements: Gemma 3’s training dataset now includes double the amount of multilingual content compared to Gemma 2.

- The same SentencePiece tokenizer (with 262K entries) is used, but it now encodes Chinese, Japanese, and Korean with improved fidelity, empowering the models to support over 140 languages for the larger variants.



	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
MGSM	18.7	57.3	68.0	2.04	34.7	64.3	**74.3**
GMMLU	43.3	64.0	69.4	24.9	57.0	69.4	**75.7**
WMT24++	38.8	50.3	53.0	36.7	48.4	53.9	**55.7**
Flores	30.2	41.3	44.3	29.5	39.2	46.0	**48.8**
XQuAD	53.7	72.2	73.9	43.9	68.0	74.5	**76.8**
ECLeKTic	8.29	14.0	17.1	4.69	11.0	17.2	**24.4**
IndicGB	47.4	59.3	62.1	41.4	57.2	61.7	**63.4**

Table 13 | Multilingual performance after the pre-training phase. IndicGenBench is an average over benchmarks reported in Table 14.

Architectural Enhancements

Gemma 3 comes with significant architectural updates that address key challenges, especially when handling long contexts and multimodal inputs. Here's what's new:

- **Optimized Attention Mechanism**: To support an extended context length of 128K tokens (with the 1B model at 32K tokens), Gemma 3 re-engineers its transformer architecture. By increasing the ratio of local to global attention layers to 5:1, the design ensures that only the global layers handle long-range dependencies while local layers operate over a shorter span (1024 tokens). This change drastically reduces the KV-cache memory overhead during inference—from a 60% increase in "global only" configurations to less than 15% with the new design.

- **Enhanced Positional Encoding**: Gemma 3 upgrades the RoPE (Rotary Positional Embedding) for global self-attention layers by increasing the base frequency from 10K to 1M while keeping it at 10K for local layers. This adjustment enables better scaling for long-context scenarios without compromising performance.

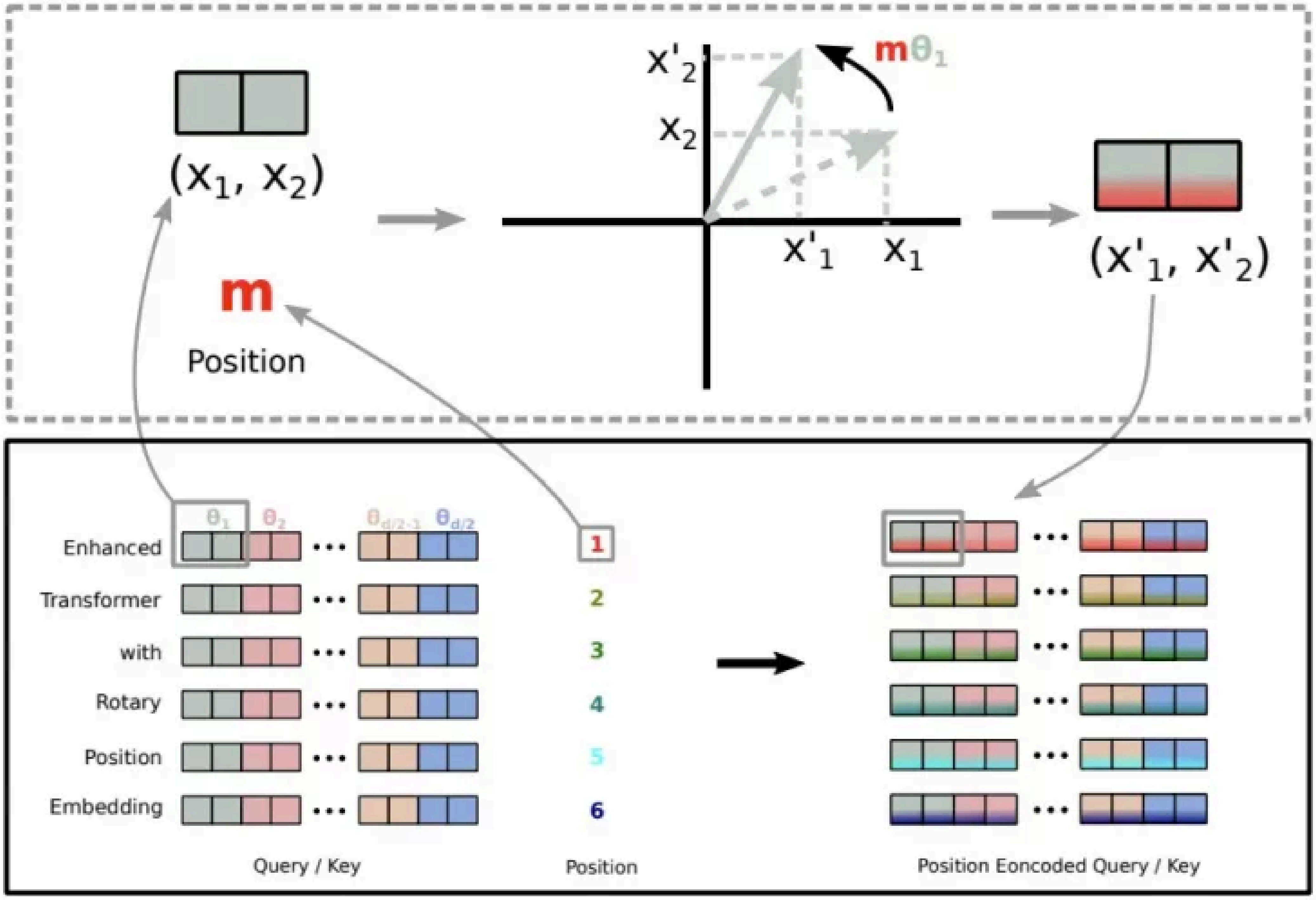


Figure 1: Implementation of Rotary Position Embedding(RoPE).

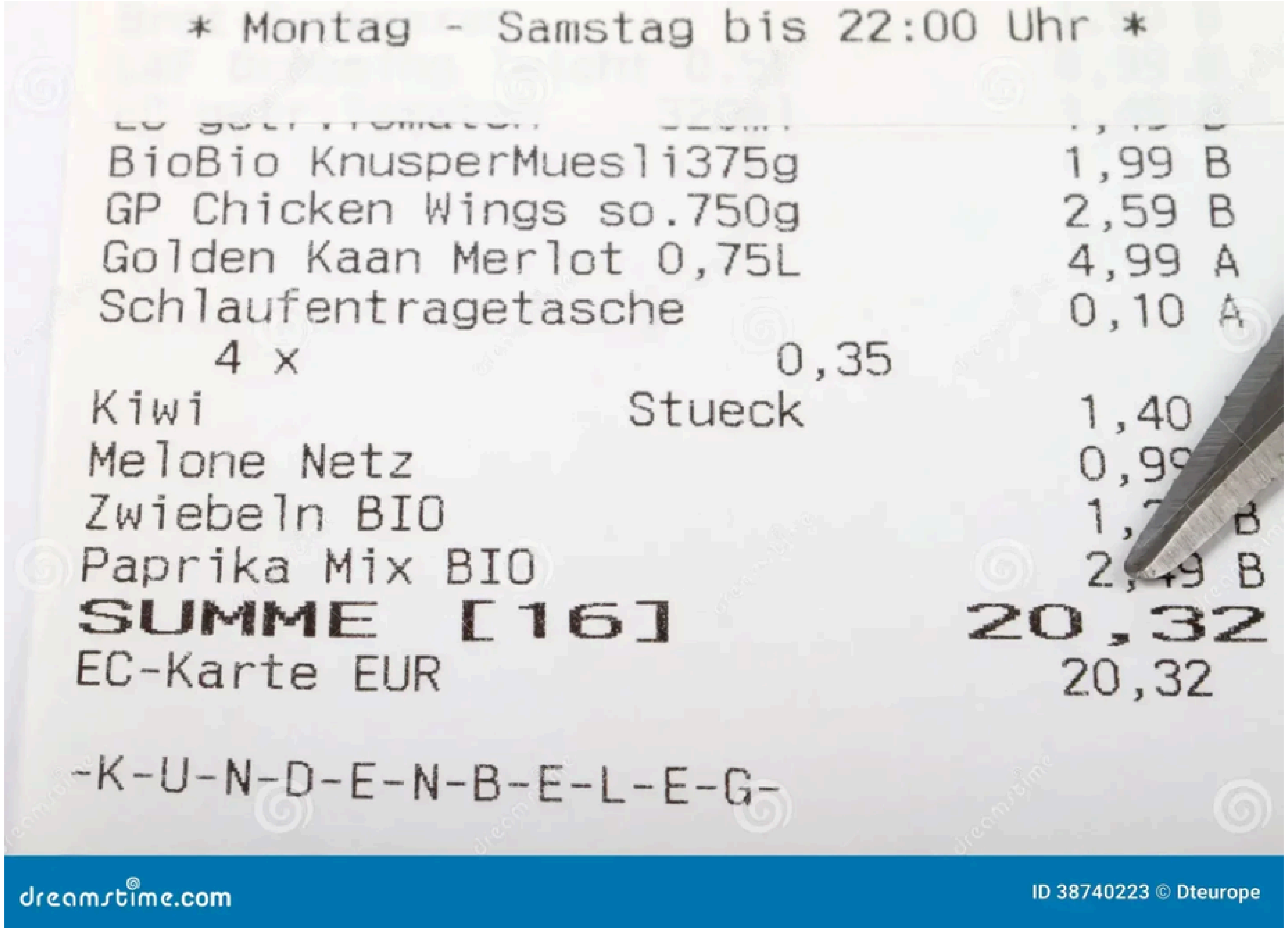
- Improved Norm Techniques: Moving beyond the soft-capping method used in Gemma 2, the new architecture incorporates QK-norm to stabilize the attention scores.

- Additionally, it utilizes Grouped-Query Attention (GQA) combined with both post-norm and pre-norm RMSNorm to ensure consistency and efficiency during training.

- **QK-Norm for Attention Scores**: Stabilizes the model's attention weights, reducing inconsistencies seen in prior iterations.

- **Grouped-Query Attention (GQA)**: Combined with both post-norm and pre-norm RMSNorm, this technique enhances training efficiency and output reliability.

- **Vision Modality Integration**: Gemma 3 expands into the multimodal arena by incorporating a vision encoder based on SigLIP. This encoder processes images as sequences of soft tokens, while a Pan & Scan (P&S) method optimizes image input by adaptively cropping and resizing non-standard aspect ratios, ensuring that the visual details remain intact.

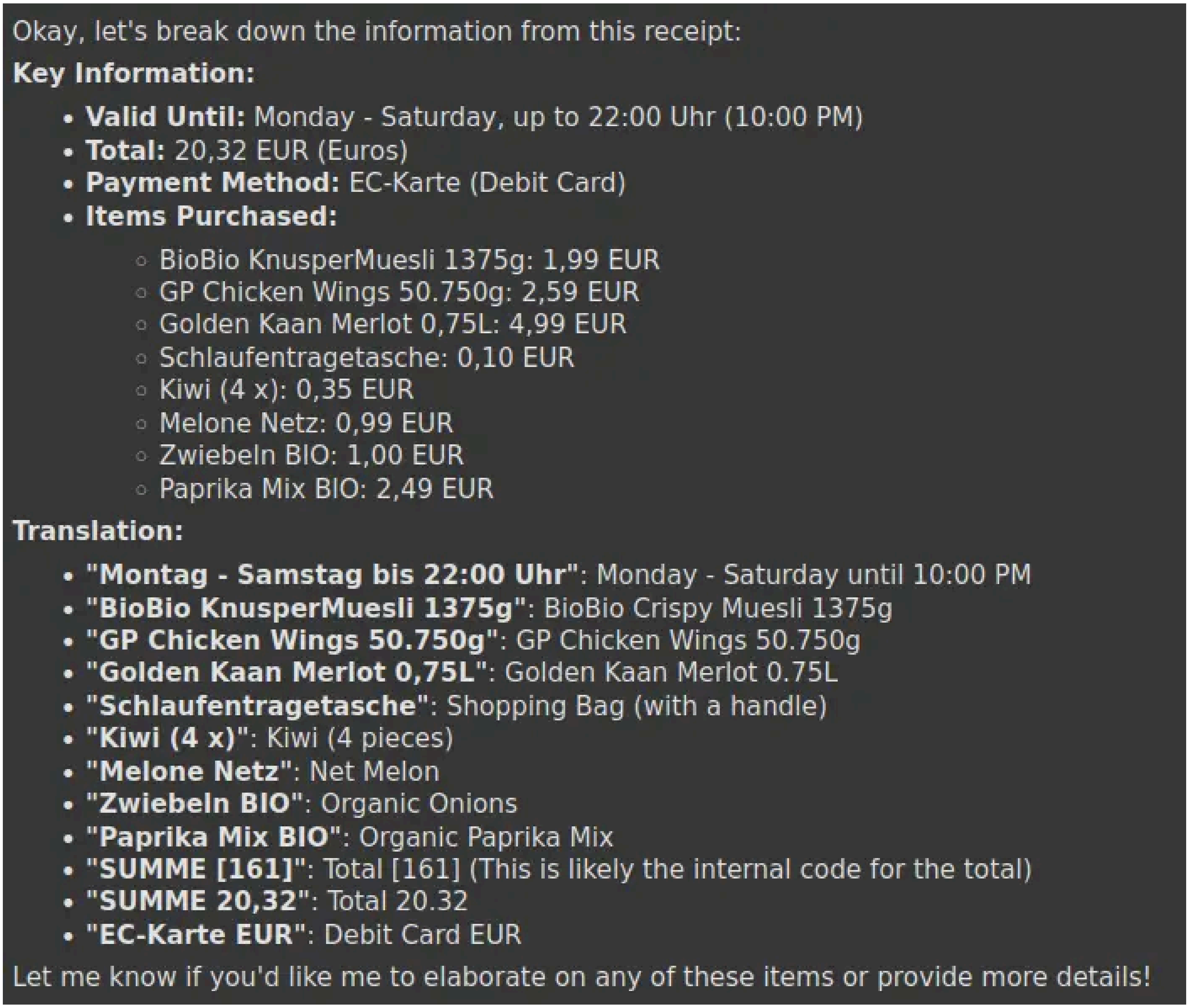


```
   * Montag - Samstag bis 22:00 Uhr *

BioBio KnusperMuesli375g          1,99 B
GP Chicken Wings so.750g          2,59 B
Golden Kaan Merlot 0,75L          4,99 A
Schlaufentragetasche              0,10 A
    4 x                  0,35
Kiwi                Stueck        1,40
Melone Netz                       0,99
Zwiebeln BIO                      1,7[illegible] B
Paprika Mix BIO                   2,49 B
SUMME  [16]                      20,32
EC-Karte EUR                      20,32

-K-U-N-D-E-N-B-E-L-E-G-
```

dreamstime.com

ID 38740223 © Dteurope

Output



Okay, let's break down the information from this receipt:

Key Information:

- **Valid Until:** Monday - Saturday, up to 22:00 Uhr (10:00 PM)
- **Total:** 20,32 EUR (Euros)
- **Payment Method:** EC-Karte (Debit Card)
- **Items Purchased:**
  - BioBio KnusperMuesli 1375g: 1,99 EUR
  - GP Chicken Wings 50.750g: 2,59 EUR
  - Golden Kaan Merlot 0,75L: 4,99 EUR
  - Schlaufentragetasche: 0,10 EUR
  - Kiwi (4 x): 0,35 EUR
  - Melone Netz: 0,99 EUR
  - Zwiebeln BIO: 1,00 EUR
  - Paprika Mix BIO: 2,49 EUR

Translation:

- **"Montag - Samstag bis 22:00 Uhr"**: Monday - Saturday until 10:00 PM
- **"BioBio KnusperMuesli 1375g"**: BioBio Crispy Muesli 1375g
- **"GP Chicken Wings 50.750g"**: GP Chicken Wings 50.750g
- **"Golden Kaan Merlot 0,75L"**: Golden Kaan Merlot 0.75L
- **"Schlaufentragetasche"**: Shopping Bag (with a handle)
- **"Kiwi (4 x)"**: Kiwi (4 pieces)
- **"Melone Netz"**: Net Melon
- **"Zwiebeln BIO"**: Organic Onions
- **"Paprika Mix BIO"**: Organic Paprika Mix
- **"SUMME [161]"**: Total [161] (This is likely the internal code for the total)
- **"SUMME 20,32"**: Total 20.32
- **"EC-Karte EUR"**: Debit Card EUR

Let me know if you'd like me to elaborate on any of these items or provide more details!

These architectural changes not only boost performance but also significantly enhance efficiency, enabling Gemma 3 to handle longer contexts and integrate image data seamlessly, all while reducing memory overhead.