EmojiSense:

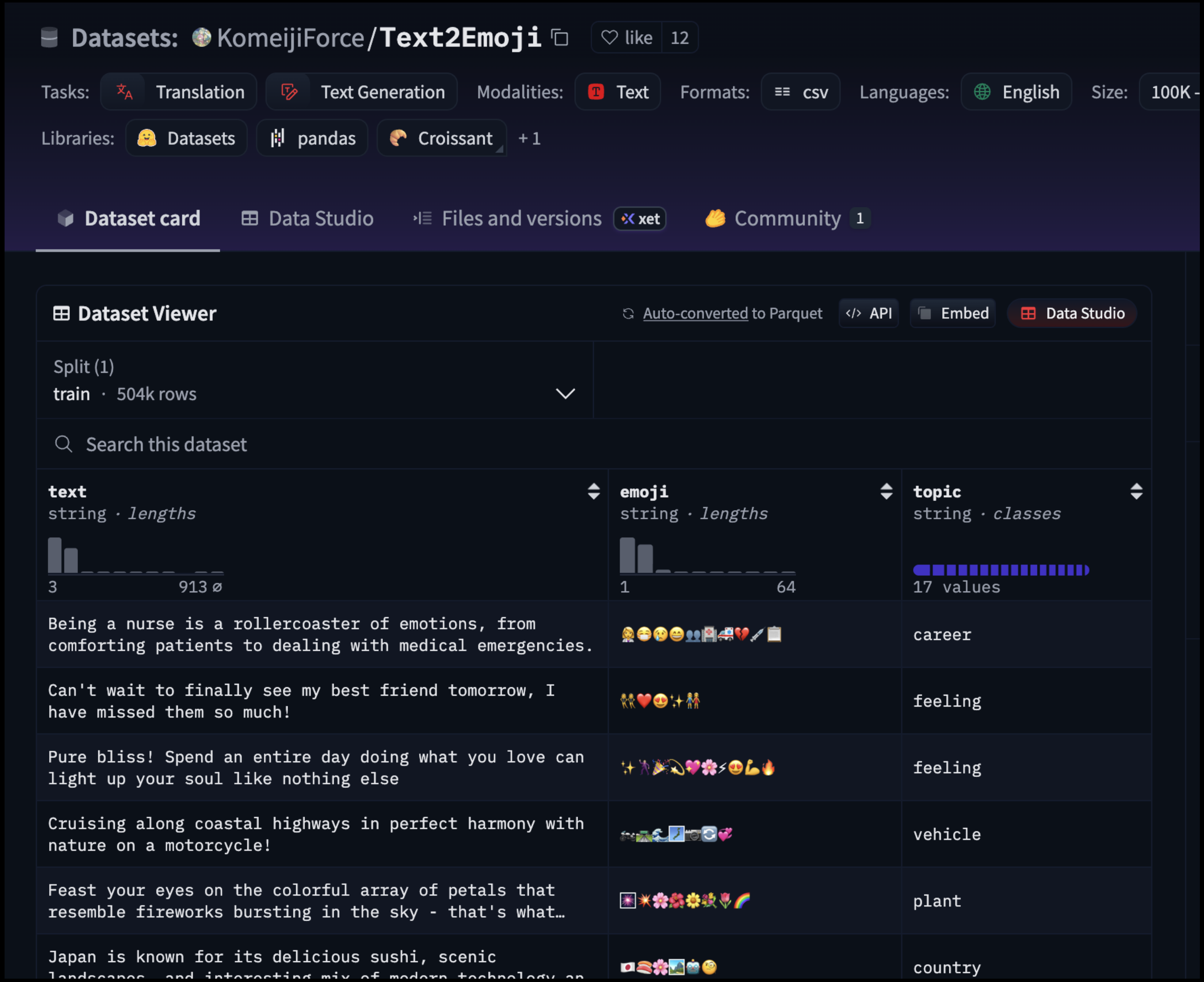
Emoji2Text Interpreter

Content

- Data & Validation
- Fine-tuning Phase
- Limitation and Future Work

Where we got our data

- We use the *Text2Emoji* dataset from hugging face (Peng et al., 2023)
- contains 504,000 rows of text to (pure) emoji data samples
- synthetically generated using ChatGPT-3.5 (LIMITATION!)



Cross-LLM Validation

```
PROMPT = """You are an expert in interpreting emojis into natural English sentences.
Here are some examples:

🔥🎉🌙 → We're having a bonfire party tonight.
📄🦢⏰ → The origami crane took forever to fold.
💐💒😊 → I caught the bouquet at the wedding.
❄️🏔️🛣️🚫 → The avalanche blocked the mountain road.

Now interpret the following emoji sequence in a similar way.
Write one fluent English sentence.

Emoji: {emoji}
Sentence:"""
```

```python
def cosine(a: str, b: str) -> float:
    va = embedder.encode(a, convert_to_tensor=True, normalize_embeddings=True)
    vb = embedder.encode(b, convert_to_tensor=True, normalize_embeddings=True)
    return float(st_util.cos_sim(va, vb).item())
```

- **Validation Model:** *Mistral* (Open Source LLM)
- **Reverse-Generation Task:** Few shot promoting with Mistral to reconstruct text sequences solely from the emoji inputs
- **Semantic Comparison:** Calculated **cosine similarity** between the *original text* and the *Mistral-generated text*
- **Filtering Threshold:** Discarded samples with similarity scores **< 0.6**
- **Final Outcome:** A curated, high-quality dataset of **200k samples**

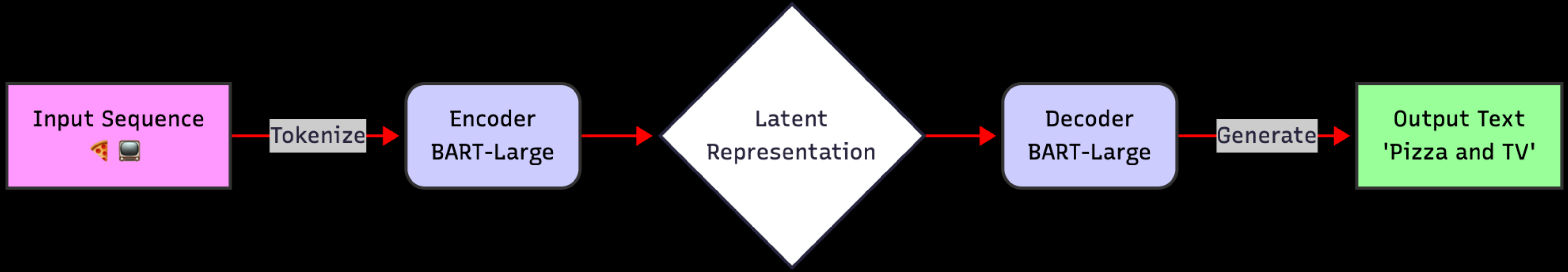
Content

- Data & Validation
- Fine-tuning Phase
- Limitations and Future Work

Step 1: Baseline model

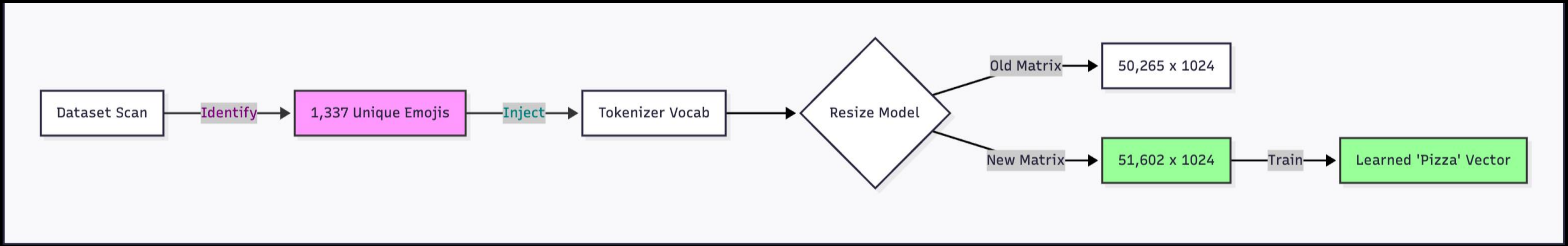
Model & Data

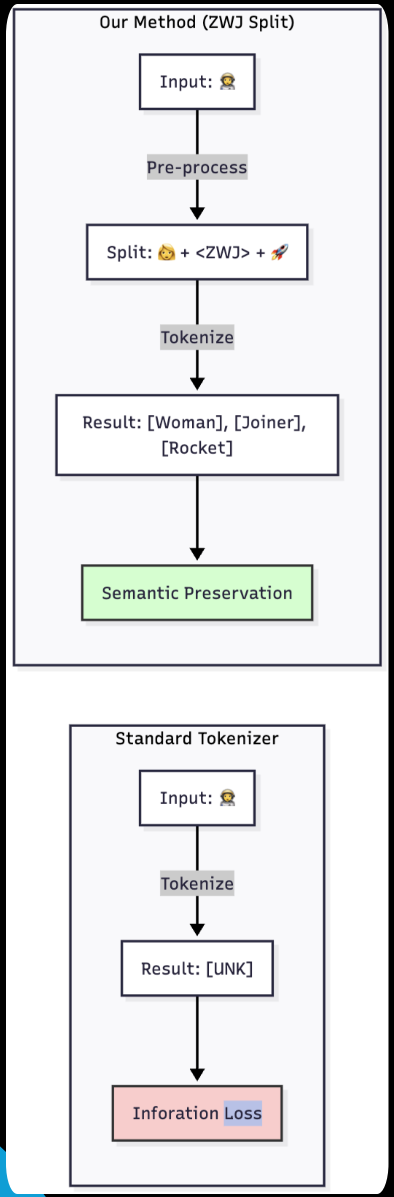
- Utilized **BART-Large** (FaceBook/bart-large), a pretrained Transformer encoder-decoder architecture optimized for translation tasks
- The Dataset: 100k samples randomly selected from the validated 200k dataset, split 80/10/10 for train/validation/test
- The Objective: Treat Emojis as a distinct source language and English as the target language (Emoji → English text)



The Tokenization Problem

- **The Limitation:** Pre-trained BART models are trained on English text (Wikipedia, Books). **They have zero emojis in their vocabulary**
- **The Consequence:** Without modification, BART sees every emoji as an <unk>
  - Input: 🍕 → Model Sees: [UNK] → Output: Random guessing
- **The Fix:**
  - Scanned the entire dataset to identify **1,337 unique emojis**
  - **Manually injected** these tokens into the BART tokenizer
  - **Resized the Embedding Layer** to learn these new emojis from scratch





The ZWJ "Grammar"

- **The Problem:** Standard tokenizers treat complex emojis (e.g., 👩‍🚀) as single, unknown tokens (<unk>)
- **The Insight:** Complex emojis are compositional. They are distinct Unicode characters glued together by a **Zero Width Joiner (ZWJ)**
- **The Solution:** Implemented a custom pre-processing function that explicitly splits emojis by the `\u200d` (ZWJ) character and use it during tokenization

Model Outputs and Human Evaluation

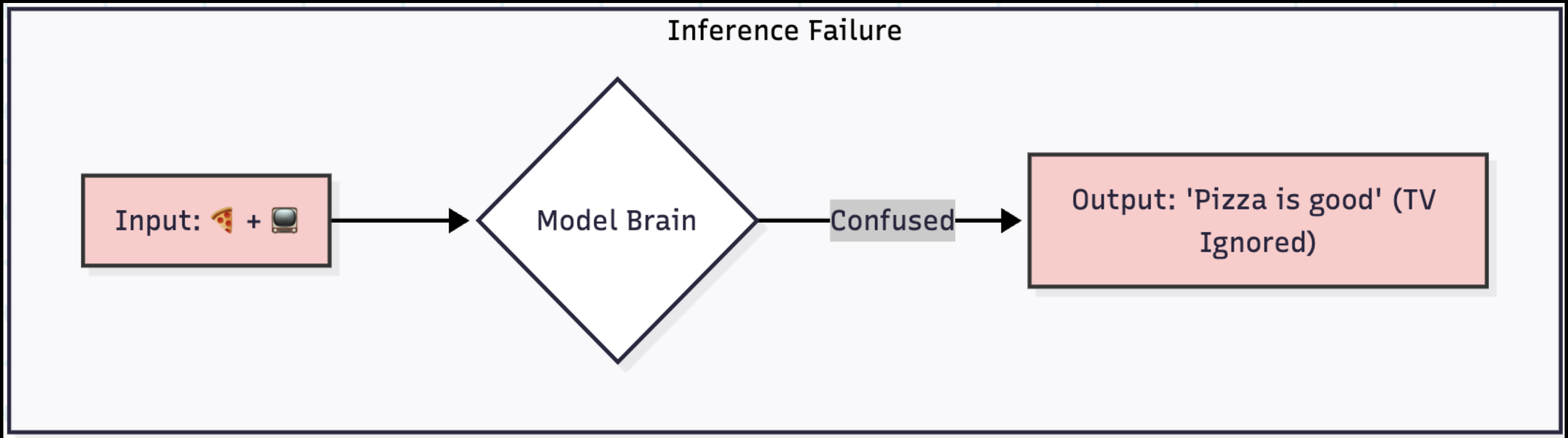
Input: 🍕 📺 🛋️
Pred : Pizza is the perfect combination of cheesy goodness and deliciousness.

Input: 🤕 🏥 💊
Prediction: I am passionate about pursuing a fulfilling career in the field of medicine.

Input: 🐍 🌿 👀
Pred : The Venus Flytrap is a carnivorous plant that captures insects with its trap-like leaves.

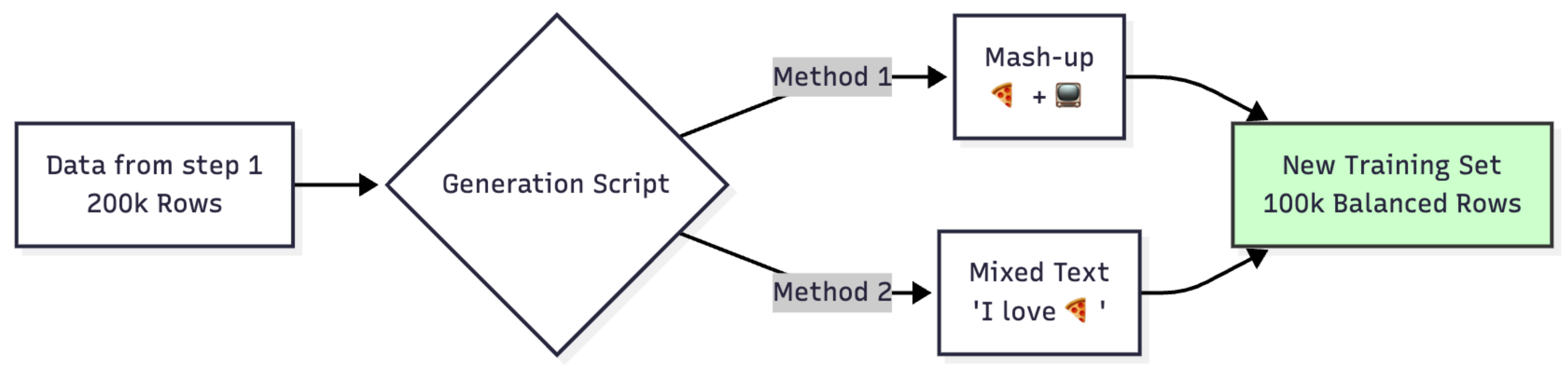
It's not that good.... Why?

Step 2: Multi-Domain & Multi-Modal Model



The "Isolated Domains" Problem

- **The Problem**: The training data (based on Emoji2Text) were "disjoint"
  - The dataset contained 19 distinct domains (Food, Activities, Weather)
  - The model never saw a "Food" emoji next to an "Activity" emoji
- Example: When faced with 🍕 📺 (Pizza + TV), the model picked the strongest signal ("Pizza") and completely ignored the other ("TV"), failing to synthesize the story



The Fix: Synthetic Data Augmentation

- **Method A: "The Mash-up" (Context Mixing)**
  - Randomly concatenated examples from different domains
- **Method B: "Mixed Modality" (Code-Switching)**
  - Randomly injected emojis into standard English sentences
  - *Example:* "I love pizza" → Input: "I love 🍕 "
- **Scale:** Generated a synthetic balanced dataset of **100,000** examples (70k Original + 15k Mashup + 15k Mixed Modality)

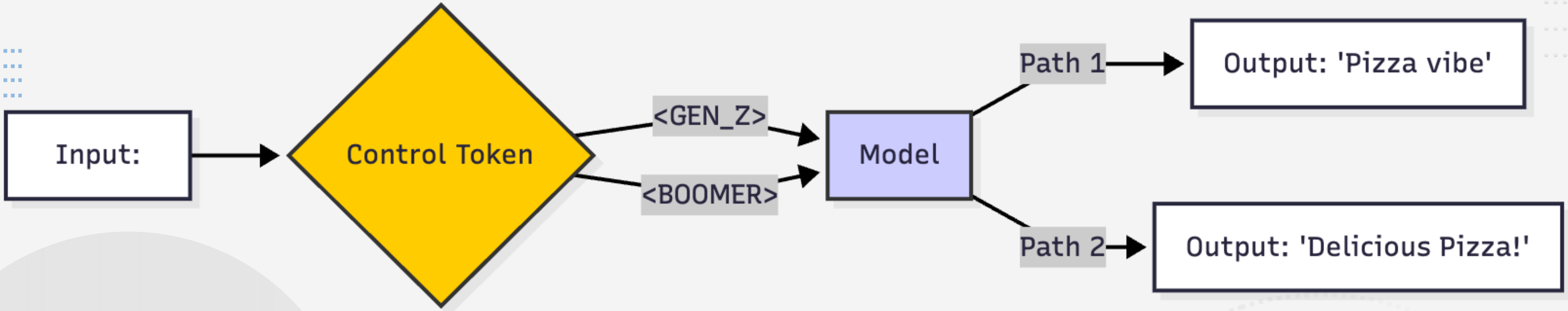
Step 2 Results and Human Evaluation

Trained a fresh BART-Large model on the new dataset for **10 epochs** using **A100 GPUs** to ensure deep convergence

Conclusion: The model is now **Context-Aware** and capable of **Multi-Modal** understanding

Input Type	Input Sequence	Baseline Output	Model 2 Output
Mash-up	🍕📺	"Pizza is delicious."	"Pizza and Netflix!"
Mixed Text	My 🚗 broke.	"My."	"My car broke."

Step 3: The Age Parameter

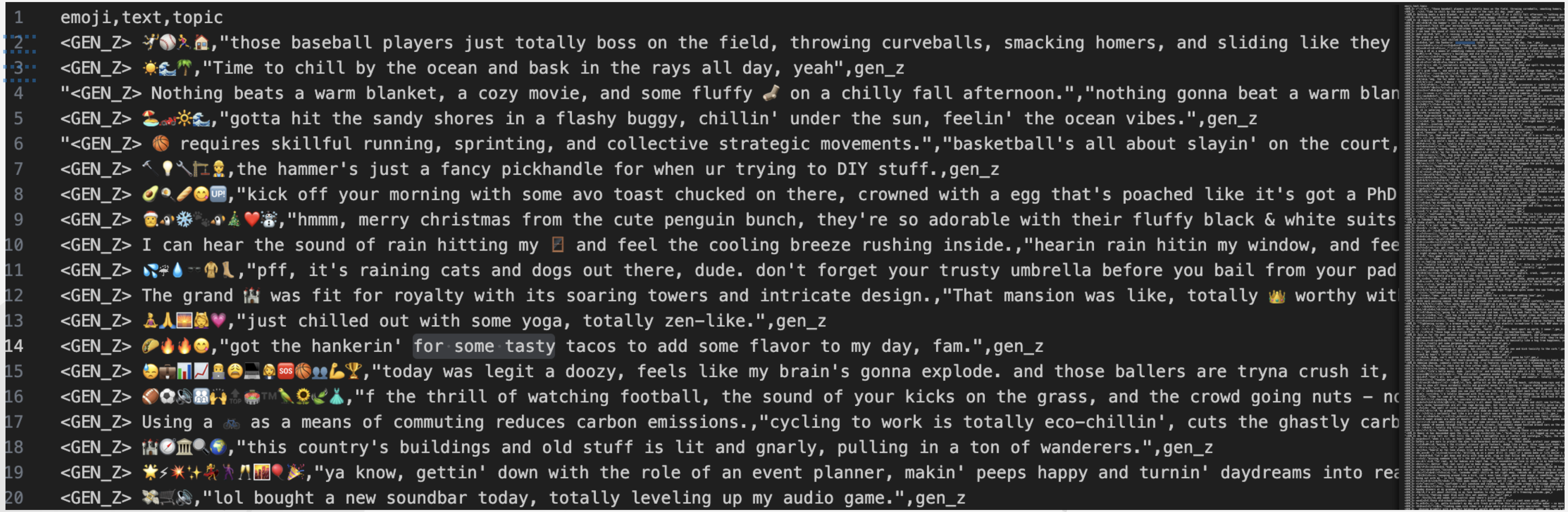


translation isn't just about meaning; it's about *tone!*

The Challenge: "Standard" English is often too robotic for emojis. A teenager (💀) and a grandparent (💀) use the same symbol to mean completely different things ("Laughter" vs. "Death")

The Solution: Introduced **Control Tokens** to append to the input and condition the model's generation

- <STD>: Standard/Neutral translation
- <GEN_Z>: Slang, lowercase, hyperbolic tone
- <BOOMER>: Formal, sincere, correct grammar & punctuation

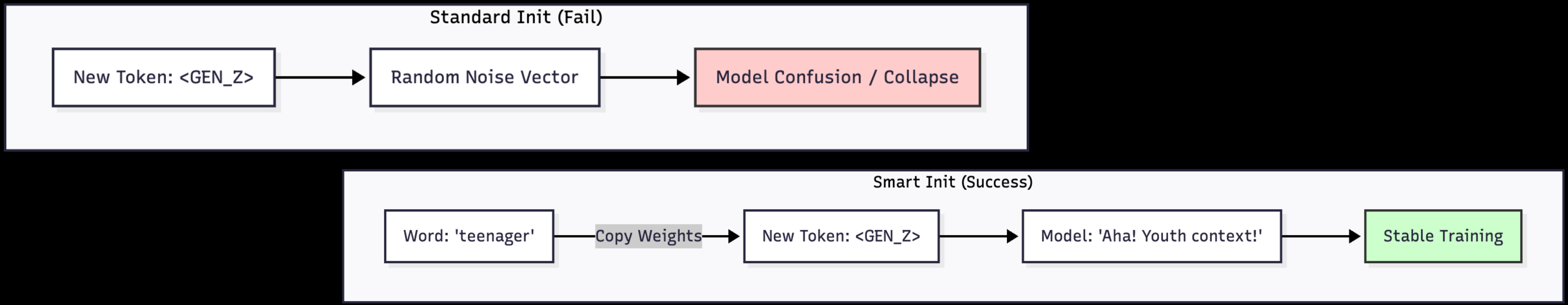


```
emoji,text,topic
<GEN_Z> 🤸⚾🏃🏠,"those baseball players just totally boss on the field, throwing curveballs, smacking homers, and sliding like they
<GEN_Z> ☀️🌊🌴,"Time to chill by the ocean and bask in the rays all day, yeah",gen_z
"<GEN_Z> Nothing beats a warm blanket, a cozy movie, and some fluffy 🧦 on a chilly fall afternoon.","nothing gonna beat a warm blan
<GEN_Z> 🏖️🏎️☀️🌊,"gotta hit the sandy shores in a flashy buggy, chillin' under the sun, feelin' the ocean vibes.",gen_z
"<GEN_Z> 🏀 requires skillful running, sprinting, and collective strategic movements.","basketball's all about slayin' on the court,
<GEN_Z> ⛏️💡🔧🏗️👷,the hammer's just a fancy pickhandle for when ur trying to DIY stuff.,gen_z
<GEN_Z> 🥑🍳🥖😋🆙,"kick off your morning with some avo toast chucked on there, crowned with a egg that's poached like it's got a PhD
<GEN_Z> 🎅🐧❄️🐾🐧🎄❤️☃️,"hmmm, merry christmas from the cute penguin bunch, they're so adorable with their fluffy black & white suits
<GEN_Z> I can hear the sound of rain hitting my 🚪 and feel the cooling breeze rushing inside.,"hearin rain hitin my window, and fee
<GEN_Z> 🌧️☔💧🐕🧥👢,"pff, it's raining cats and dogs out there, dude. don't forget your trusty umbrella before you bail from your pad
<GEN_Z> The grand 🏰 was fit for royalty with its soaring towers and intricate design.,"That mansion was like, totally 👑 worthy wit
<GEN_Z> 🧘🙏🌅🧘💖,"just chilled out with some yoga, totally zen-like.",gen_z
<GEN_Z> 🌮🔥🔥😋,"got the hankerin' for some tasty tacos to add some flavor to my day, fam.",gen_z
<GEN_Z> 😓💼📊📈👨‍💼😩💻👩‍💼🆘🏀👥💪🏆,"today was legit a doozy, feels like my brain's gonna explode. and those ballers are tryna crush it,
<GEN_Z> 🏈⚽🔊👨‍👩‍👧‍👦🙌🔝🏟️™️🦎🌻🌿👗,"f the thrill of watching football, the sound of your kicks on the grass, and the crowd going nuts – no
<GEN_Z> Using a 🚲 as a means of commuting reduces carbon emissions.,"cycling to work is totally eco-chillin', cuts the ghastly carb
<GEN_Z> 🏰🧭🏛️🔍🌍,"this country's buildings and old stuff is lit and gnarly, pulling in a ton of wanderers.",gen_z
<GEN_Z> 🌟⚡💥✨💃🕺🥂🌆🎈🎉,"ya know, gettin' down with the role of an event planner, makin' peeps happy and turnin' daydreams into rea
<GEN_Z> 🎶🛒🔊,"lol bought a new soundbar today, totally leveling up my audio game.",gen_z
```

The "Teacher-Student" Loop

- **Synthetic Data Generation:** We used a "Teacher" LLM (*Microsoft Phi-3*) to generate training data for our "Student" model (BART)
- **Prompt Engineering:** Rewrote Step 2 sentences into specific personas
    - `"You are a Gen Z translator. Rewrite the input sentence using slang, lowercase, and irony."`
- **Dataset Composition:** Created 40k stylistic examples (20k Gen Z, 20k Boomer) mixed with 60k Standard examples

The Collapse & The Fix



- **The Failure:** Initially, training failed. The model "collapsed," ignoring inputs and outputting repetitive garbage ("The The The...")
- **The Root Cause:** The new control tokens (<GEN_Z>) were initialized as **random noise**. This "shocked" the pre-trained model, destroying its weights during early training
- **The Fix: Smart Initialization**
  - We manually copied the embedding weights from semantically similar English words to initialize the new tokens
  - <GEN_Z> → initialized from **"teenager"**
  - <BOOMER> → initialized from **"formal"**
- **Result:** The model started training with a "hint" about what the tokens meant, leading to stable convergence

FINAL RESULTS

- **Qualitative Success:** The model successfully nuances meaning based on the persona
  - *Context Awareness:* It interprets 🚗 as "whip/car" for Gen Z and "automobile" for Boomers
  - *Tone Shift:* It changes punctuation and capitalization patterns dynamically
- **Final Deliverable:** A functional, interactive web demo running the live model

```
🤖 EMOJI-LM AGE PARAMETER EVALUATION
============================================================

📝 Input: 🍕📺 (Pizza + TV)
------------------------------
   STD      : Pizza night with friends is always a good idea!
   GEN_Z    : i'm totally vibin' with some pizza and binge-watching my fave show.
   BOOMER   : Nothing compares to the sheer delight of savoring a delectable slice of pizza on a Friday evening!
```

Content

- Data & Validation
- Fine-tuning face
- Limitations and Future Work

Limitation & Constraints

- We can't promise that we have **all** the emojis in our model's vocab...
  - Results in Out-of-Vocabulary (OOV) errors or unpredictable inference on rare emojis
- The "Teacher-Student" training loop relies entirely on the Teacher LLM's capabilities...
  - If the Teacher LLM hallucinates or exhibits bias, these flaws are distilled into our model

Possible next step

Scaling & Generalization

- Expand model capacity by increasing the training corpus size
- Extend training epochs to improve convergence and reducing loss

Further Validation

- Develop a secondary validation layer for the "Teacher-Student" pipeline (Step 3)
- **Goal:** Systematically filter low-quality or hallucinated outputs from the Teacher model to reduce noise

References

Mistral AI. 2025. Models — getting started (documentation). https://docs.mistral.ai/getting-started/models. Accessed: 2025-10-28.

Lars Henning Klein, Roland Aydin, and Robert West. 2024. Emojinize: Enriching any text with emoji translations. ArXiv:2403.03857.

Tom Krantz and Alexandra Jonker. 2024. What is cosine similarity? https://www.ibm.com/think/topics/cosine-similarity. Accessed: 2025-10-28.

Google LLC. 2024. mt5-small — multilingual text-to-text transformer model (hugging face model card). https://huggingface.co/google/mt5-small. Accessed: 2025-10-28.

Letian Peng, Zilong Wang, Hang Liu, Zihan Wang, and Jingbo Shang. 2023. Emojilm: Modeling the new emoji language. ArXiv:2311.01751.

Wikipedia. Bleu – bilingual evaluation understudy (wikipedia). https://en.wikipedia.org/wiki/BLEU. Accessed: 2025-10-28.