From language to vision

23 / 10 / 2024

Istvan Fehervari
Director, Data & ML

BenchSci

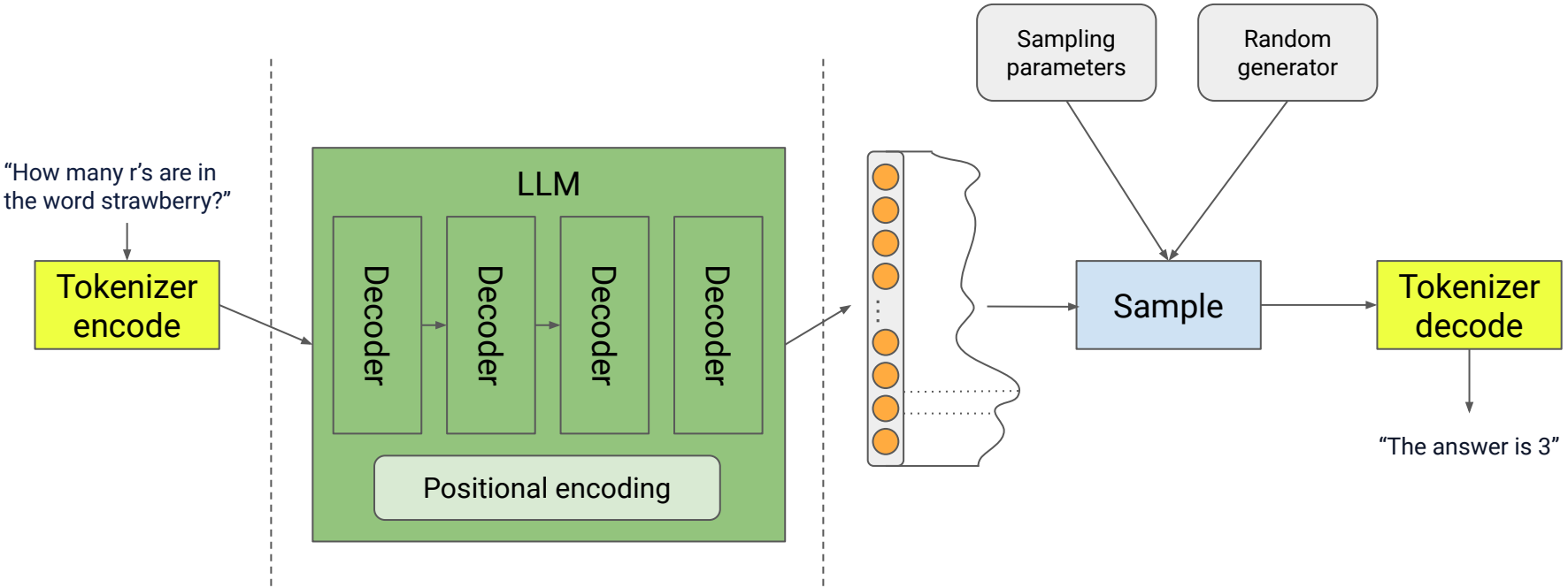
Agenda

01 What are VLMs and how do they work?

02 How do we build with VLMs?

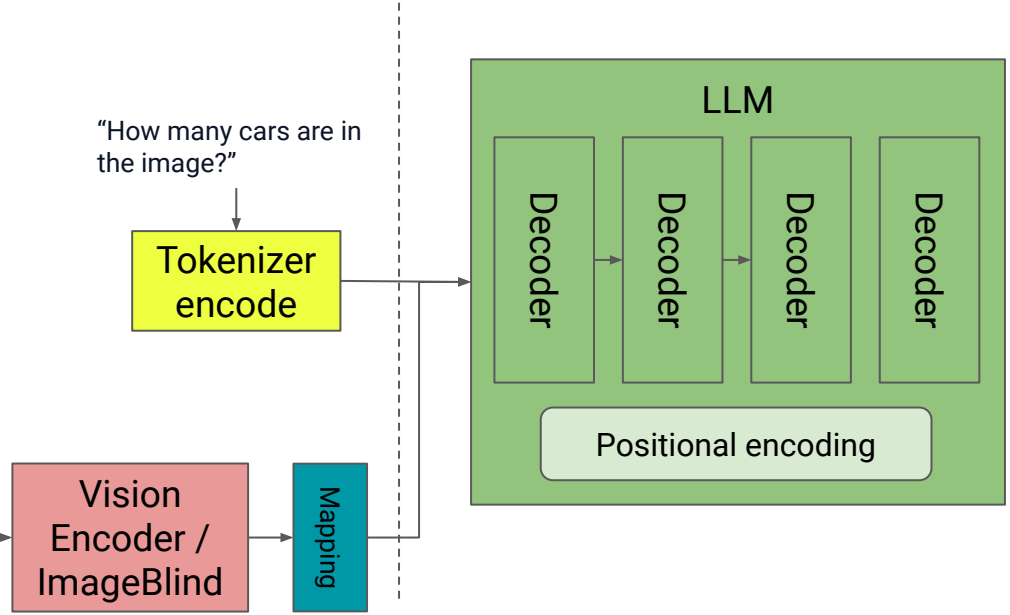
03 Discussion and Q&A

Primer on LLMs



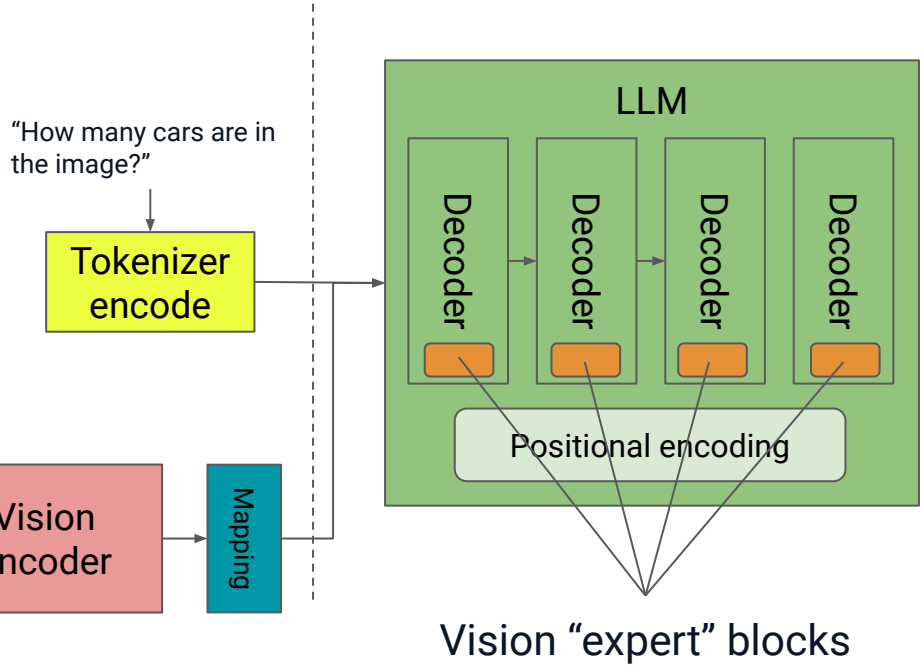
Enabling Multiple modalities

- Shallow alignment
- Vision encoder can be trained or frozen
- LLM is pre-trained and frozen
- LlaVa, PandaGPT



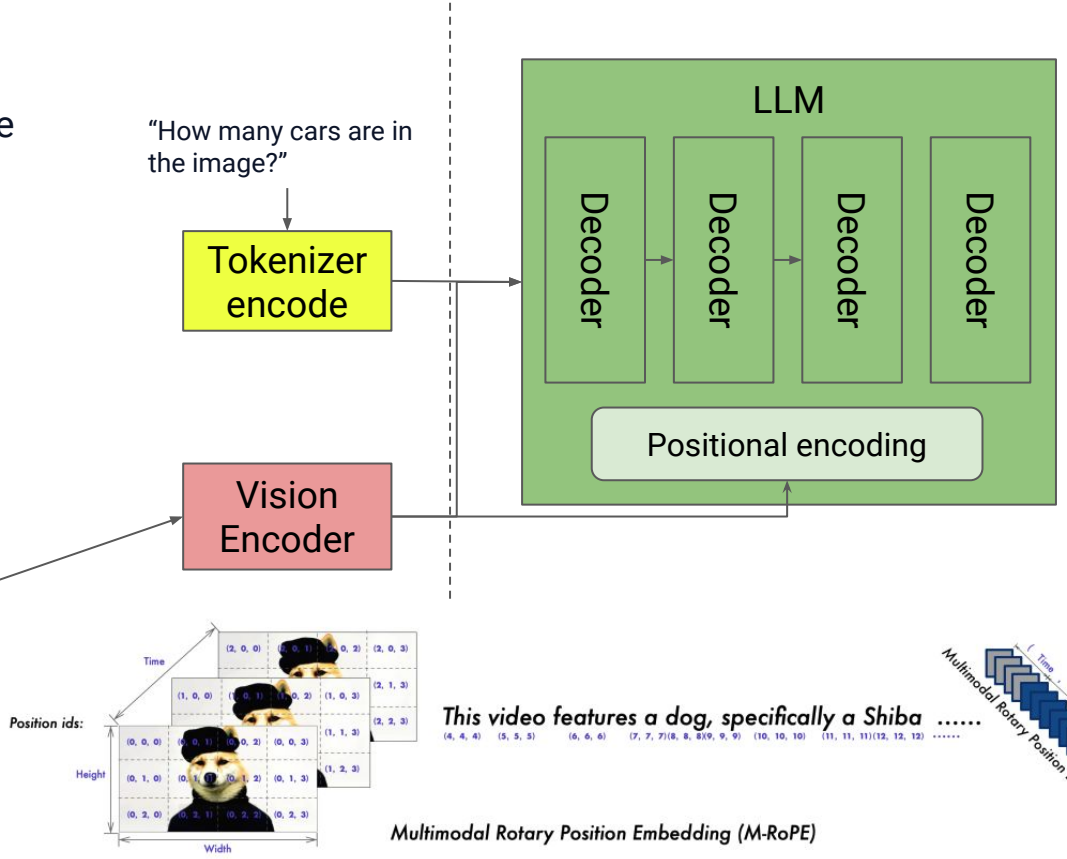
Enabling Multiple modalities

- Deep alignment
- Mixture of experts
- LLM is fine-tuned
- E.g. CogVLM



Enabling videos and any-sized images

- Vision encoders often down-sample to 224x224
- Arbitrary input resolution via 2d rotary position embedding
- E.g. Qwen2-VL (Wang et. al. 2024)



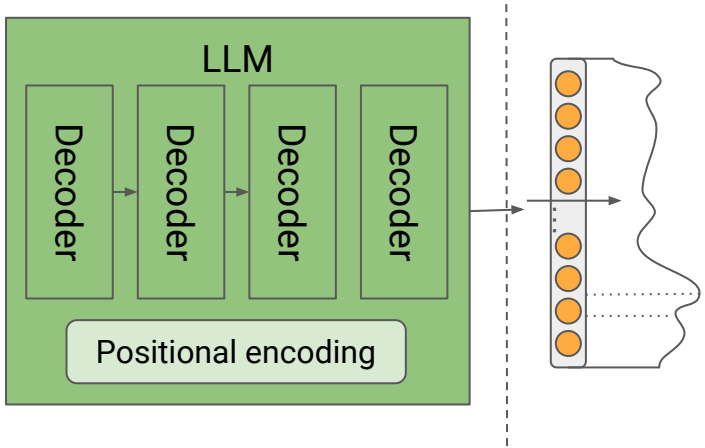
Can VLMs do perception (well)?

Great at:

- Zero/few-shot classification
- Image captioning
- Reasoning under uncertainty
  - In context learning with images
  - e.g., categorize product review based on attached image(s), robot navigation, chart understanding

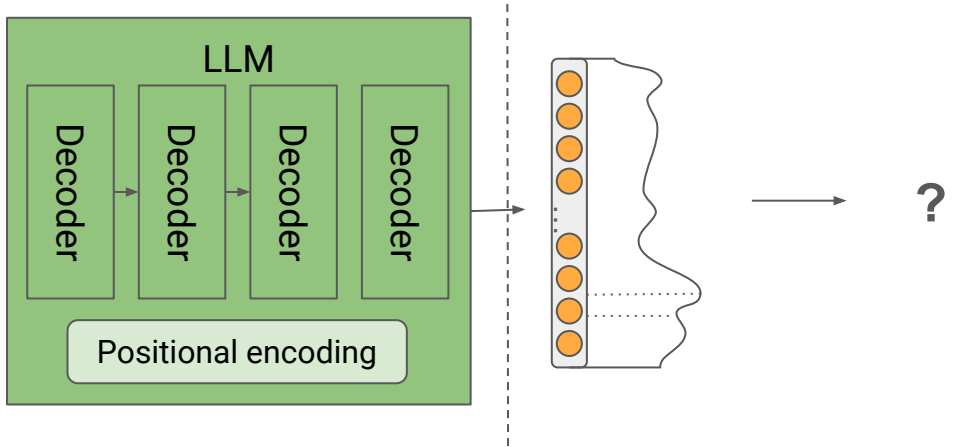
Not great at (yet):

- object counting
- Object arrangements
- predicting bounding boxes
- predicting segmentation masks



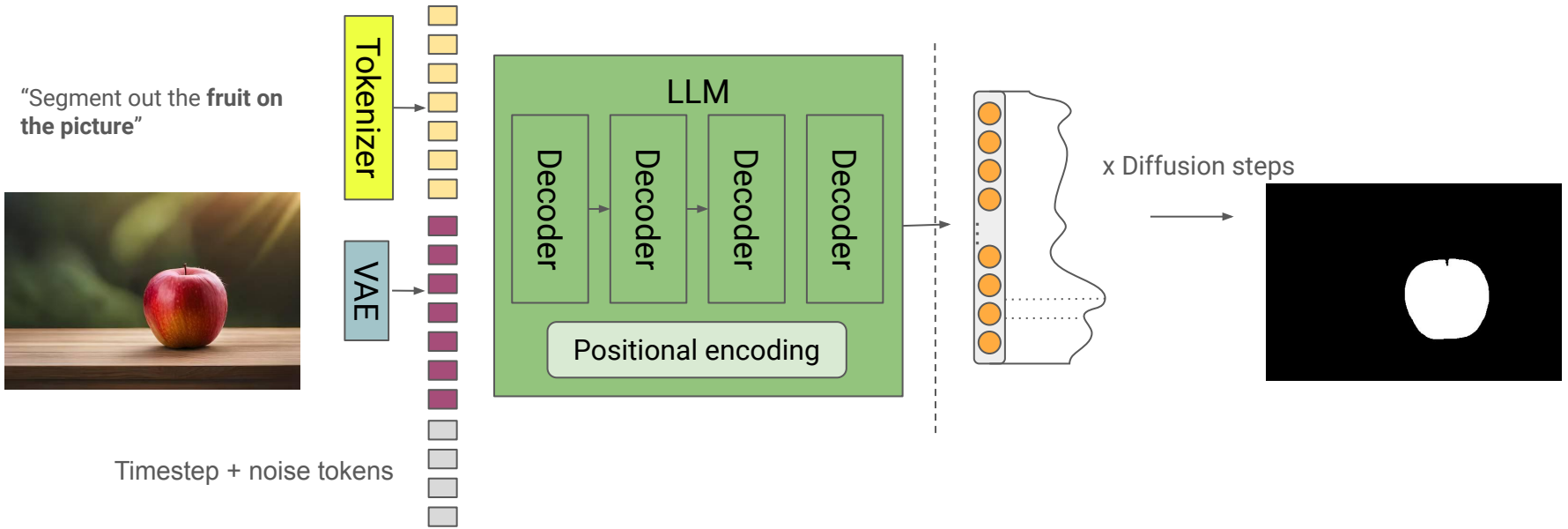
Controlling VLM predictions

- Control output generation during prompting and decoding -> automate evaluation
  - Logit biasing
    - Instructing for a few very specific tokens as outputs and using their logits to establish class-level confidence
  - Structured output
    - Pydantic / JSON / XML / YAML structures

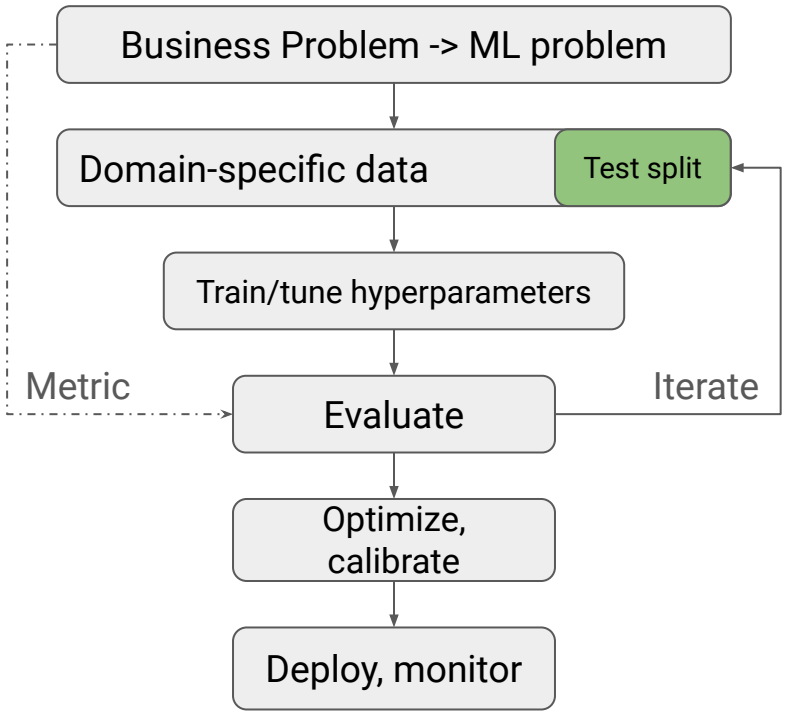


Universal modelling

- What's next: enable 2D structured outputs
- Image/mask generation with rectified flow (Xiao et. al. OmniGen, 2024)
- LLMs can compose and "reason"



Building in the pre-genai era



- Purpose-built models
- Feasibility is mostly guaranteed
- Solution quality is defined via **data**
- More data won't make things worse
- Iterations provide insights
- Interpretability is possible

Why shall I consider a VLM?

01 Need a generative solution, i.e. a piece of text

E.g. Captioning, image generation

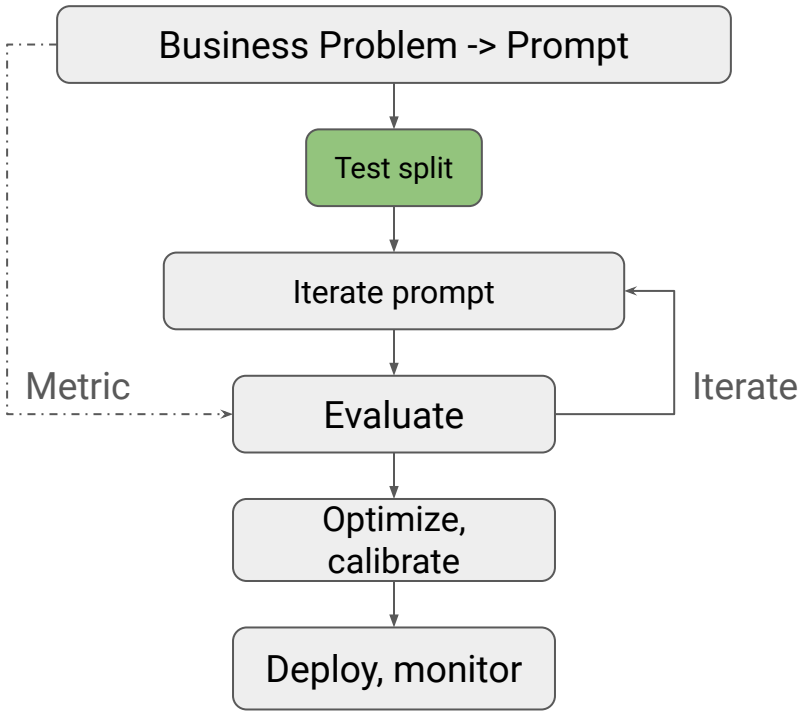
02 Problem cannot be captured via data efficiently

- High ambiguity - requires reasoning
- Context matters and/or changes frequently
- Training data does not exist / too expensive to collect

03 Need a quick solution

- Validate product viability

Building with Generative AI



- Data sampling is critical
- Feasibility is not guaranteed
- Metric definition can be challenging
- Iteration is fast
- Iteration isn’t well guided
- Easier to adapt or deploy solutions

Choosing the right model

Cloud hosted VLM

Pros:

- Best (initial) performance
- Likely cheaper per-request
- No deployment upkeep

Cons:

- **Content filtering false positives**
- Not 100% reproducible
- Need downtime mitigation strategy
- Limited decoding support

Self-hosted VLM

Pros:

- Full control over model size
- Outputs are fully reproducible
- Any decoding schema can be used
- Uptime is self-managed

Cons:

- **Libraries are catch-up**
- Likely more expensive per-request
- Lower (initial) performance
- Needs ops work to maintain

Discussion and Q&A