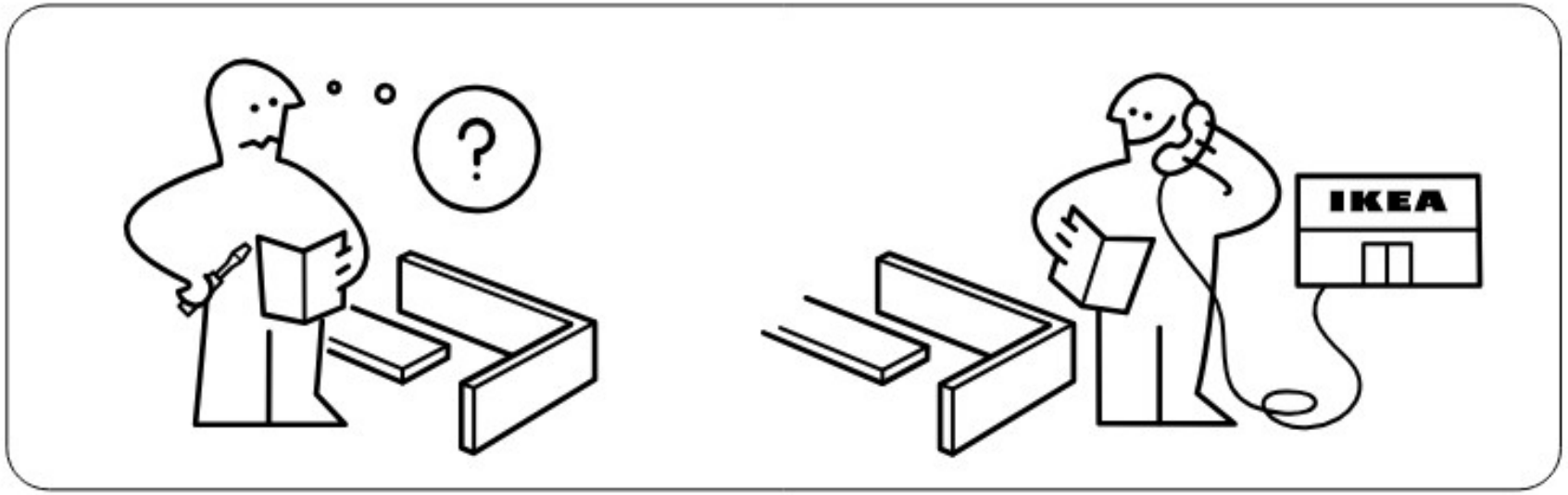
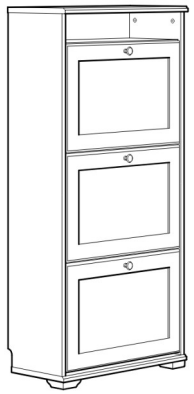
Ikea Assembly Guide Assistant



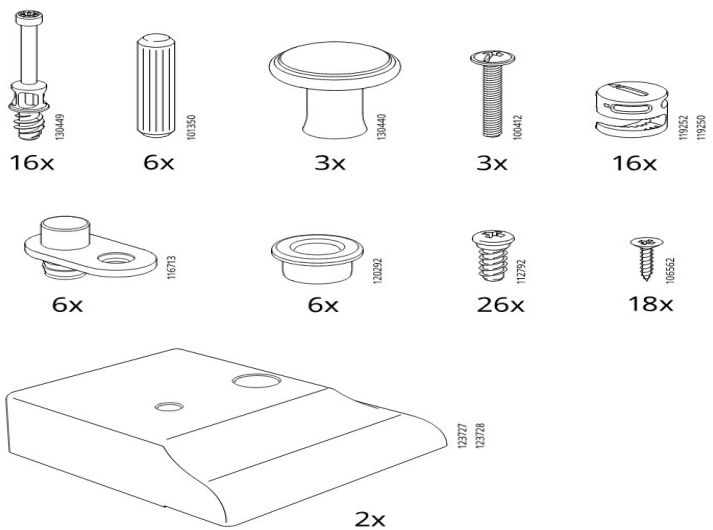
Finn, Christian, Daulet, Kenan

Systems and Software Engineering 2025

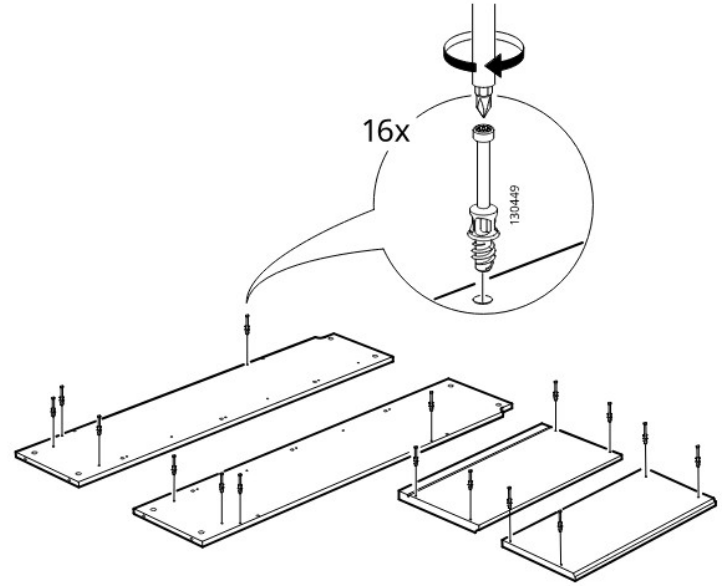
BRUSALI



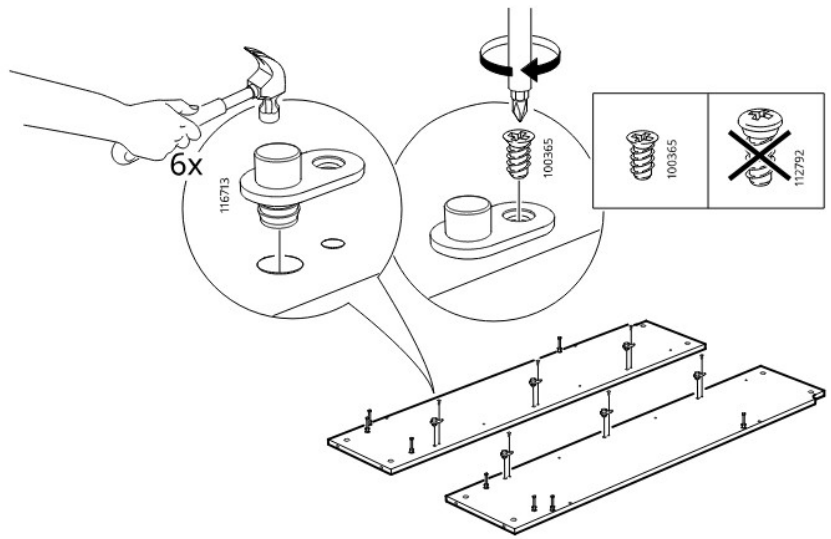
Design and Quality
IKEA of Sweden

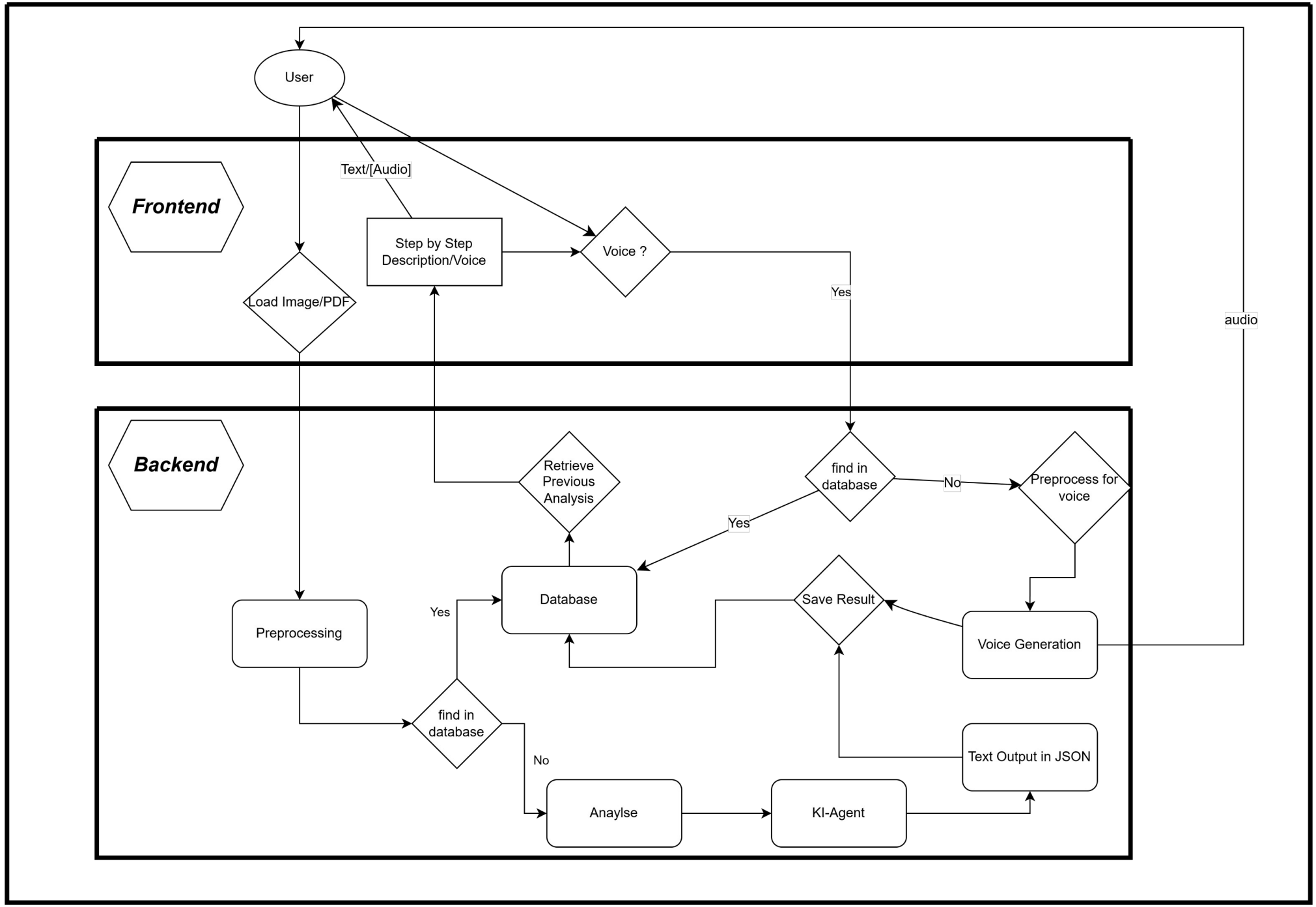


1



2

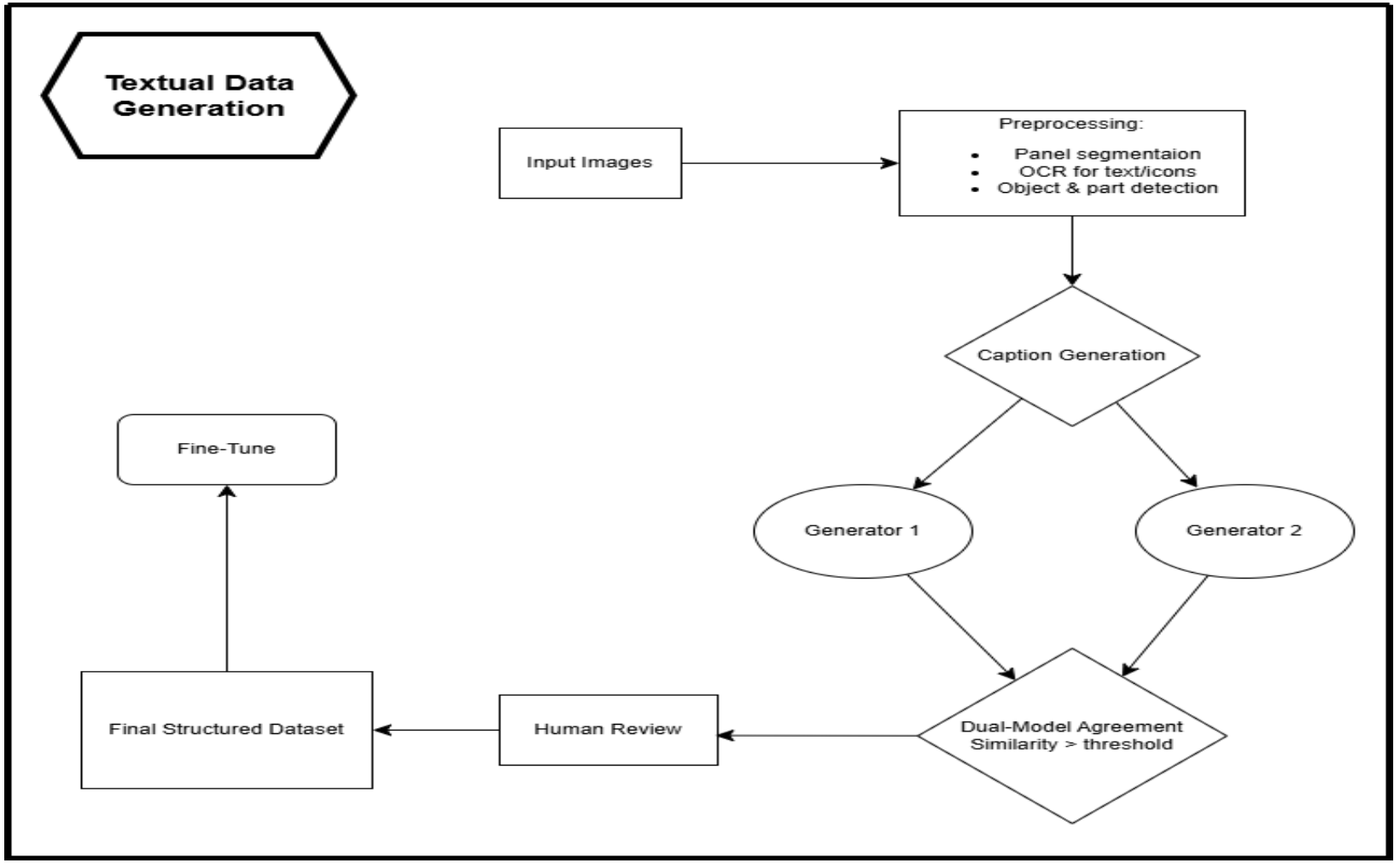



User
Frontend
Text/[Audio]
Step by Step Description/Voice
Voice ?
Load Image/PDF
Yes
audio
Backend
Retrieve Previous Analysis
find in database
No
Preprocess for voice
Yes
Database
Save Result
Preprocessing
Yes
Voice Generation
find in database
No
Text Output in JSON
Anaylse
KI-Agent

Implementation Strategies for the AI Agent

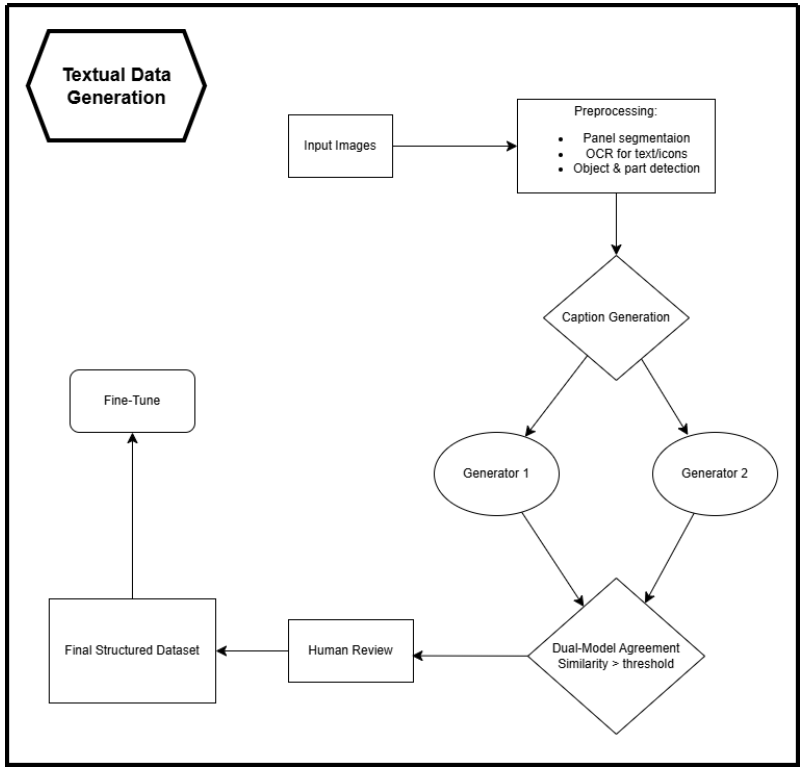
- **From Scratch**
  - Build architecture + dataset
  - Extremely resource-intensive (compute + data)
  - Not feasible right now
- **Fine-Tuning a Pretrained Model**
  - Adapt open models (e.g. CLIP, BLIP, LlaVA)
  - Domain-specific data needed (IKEA image text pairs)
  - Cheeper but performance < GPT-Vision tier
  - Promising if we generate a dataset…
- **External APIs (GPT-4V, Gemini Vision etc.)**
  - High accuracy with minimal setup
  - „Black box“ behaviour + external dependency
  - Best short-term trade-off for prototype

From Images to Dataset



From Images to Dataset

- **Image Panels → Preprocessing**
  - Detect panels, extract OCR text, identify icons & parts
- **Dual Caption Generation**
  - Use API model (e.g., GPT-4V) and open VLM (e.g., BLIP/LLaVA)
  - Generate JSON-formatted step descriptions
- **Agreement & Grounding Filter**
  - Keep samples where both models agree semantically
  - Verify that mentioned parts/tools appear in image or OCR
- **Human Verification & Linting**
  - Manually check ~5–10 % for clarity, safety notes, and consistency
  - Apply automated style rules (vocabulary, tone, structure)
- **Final Output:**
  - High-quality instruction dataset ready for fine-tuning or evaluation



Model Candidates for Fine-Tuning

- **Goal**: adapt a pretrained Vision-Language Model (VLM) to understand IKEA assembly images and generate structured instructions.

  Possible Base Models

- **CLIP from OpenAI**
  - Strong visual-text alignment
  - Good for feature extraction and image-text embedding
  - Not generative → useful as encoder backbone

Model Candidates for Fine-Tuning

- **Goal**: adapt a pretrained Vision-Language Model (VLM) to understand IKEA assembly images and generate structured instructions.

  Possible Base Models

- **BLIP / BLIP-2 (Salesforce Research)**
  - Vision encoder + language decoder
  - Supports caption generation
  - Well-suited for instruction-style text generation

Model Candidates for Fine-Tuning

- **Goal**: adapt a pretrained Vision-Language Model (VLM) to understand IKEA assembly images and generate structured instructions.

  Possible Base Models

- **LLaVA (Large Language and Vision Assistant)**
  - Combines CLIP visual encoder with LLaMA language model
  - Fine-tuning framework already available
  - Ideal for instruction-following and visual reasoning

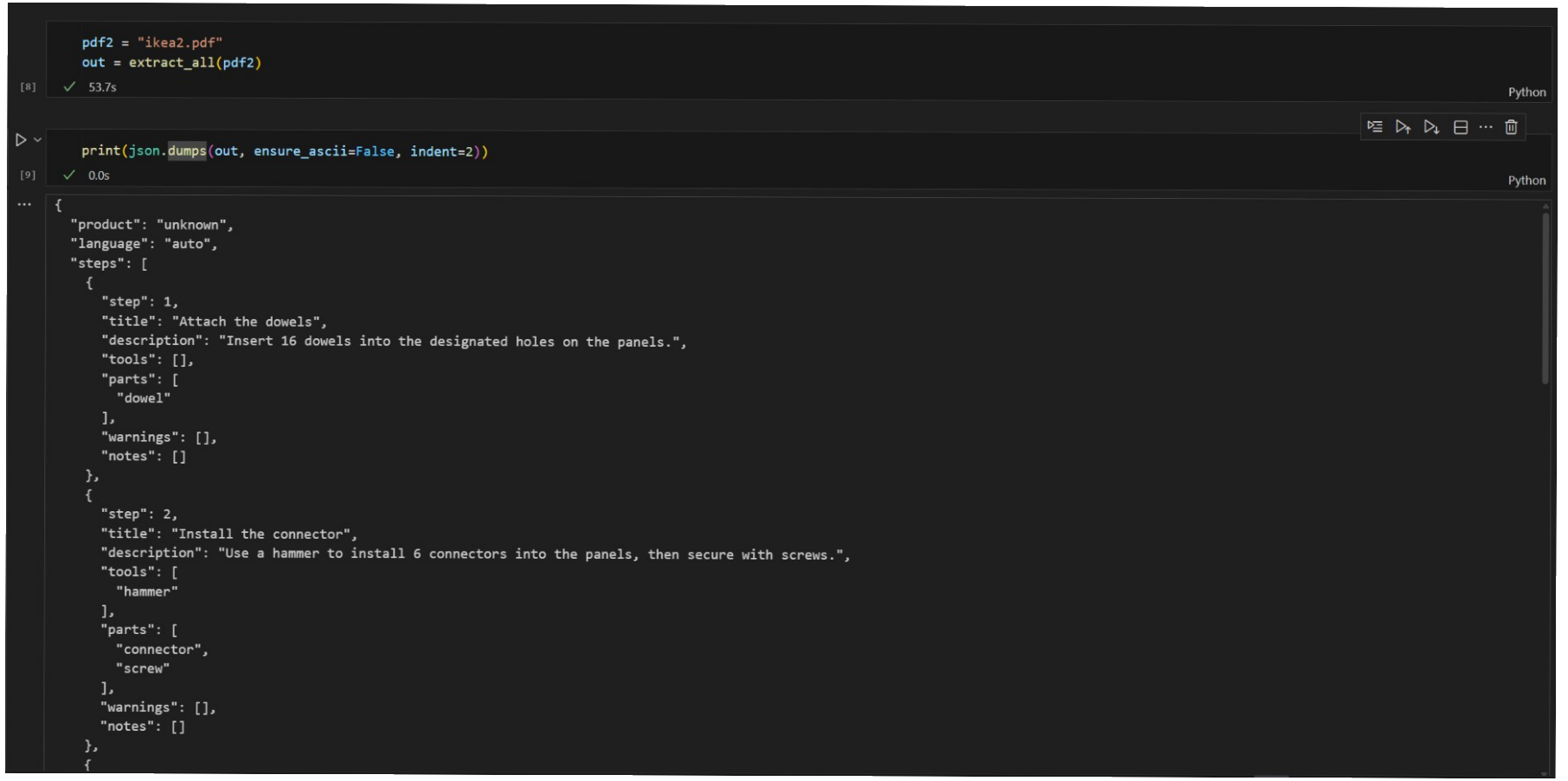
Model Candidates for Fine-Tuning

- **Goal**: adapt a pretrained Vision-Language Model (VLM) to understand IKEA assembly images and generate structured instructions.

  Possible Base Models

- **Kosmos-2 / MiniGPT-4 / InternVL-2**
  - Multimodal models that can reason about text and visuals
  - Easier to fine-tune with LoRA (fine-tune large models efficiently, train a very small numer of new parameters, while keeping the original model frozen.)
  - Compatible with image–text pairs like IKEA steps

Example: GPT Vision API for text generation



```
pdf2 = "ikea2.pdf"
out = extract_all(pdf2)
```

[8] ✓ 53.7s Python

```
print(json.dumps(out, ensure_ascii=False, indent=2))
```

[9] ✓ 0.0s Python

```
{
  "product": "unknown",
  "language": "auto",
  "steps": [
    {
      "step": 1,
      "title": "Attach the dowels",
      "description": "Insert 16 dowels into the designated holes on the panels.",
      "tools": [],
      "parts": [
        "dowel"
      ],
      "warnings": [],
      "notes": []
    },
    {
      "step": 2,
      "title": "Install the connector",
      "description": "Use a hammer to install 6 connectors into the panels, then secure with screws.",
      "tools": [
        "hammer"
      ],
      "parts": [
        "connector",
        "screw"
      ],
      "warnings": [],
      "notes": []
    },
    {
```