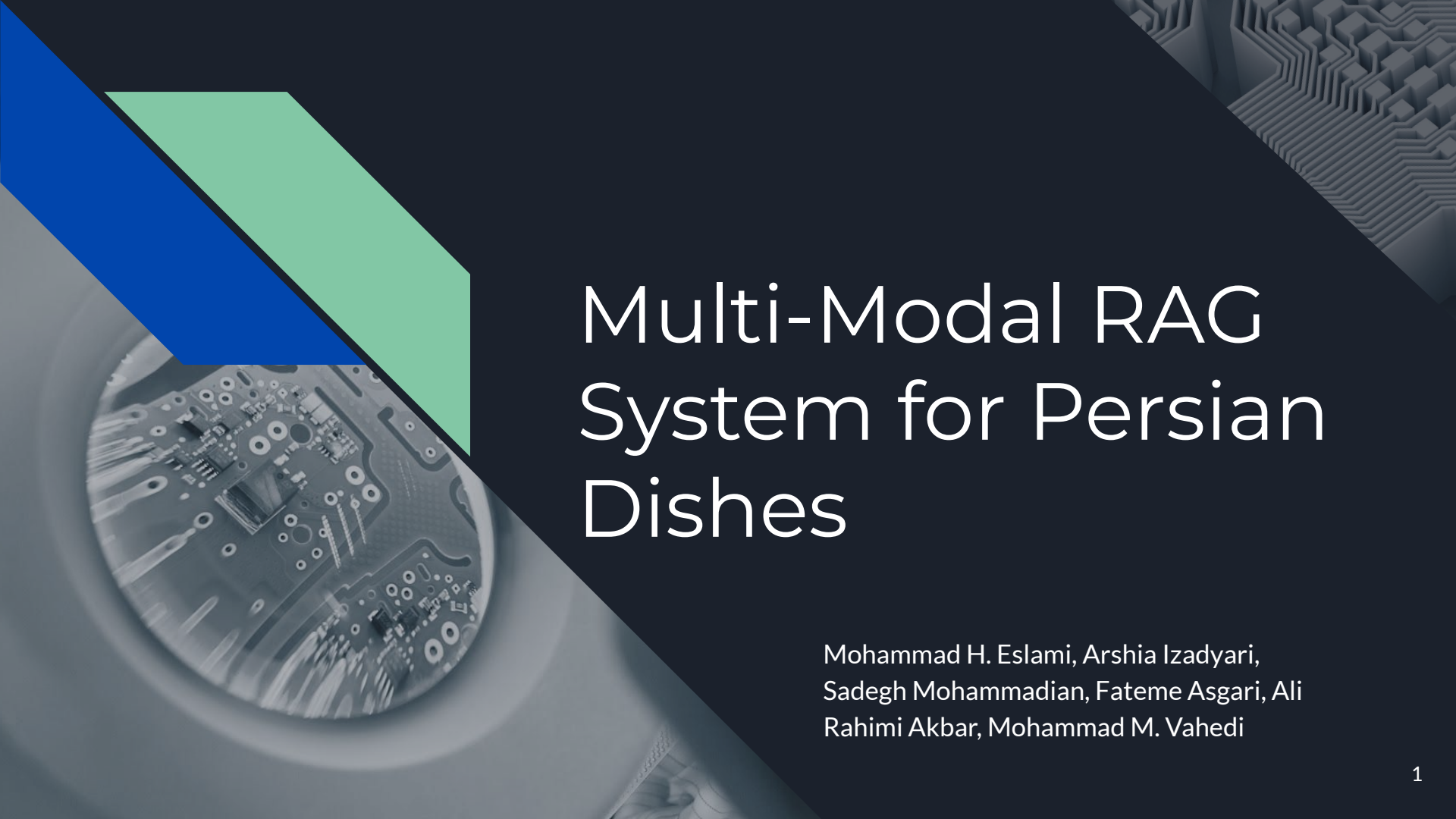


Multi-Modal RAG System for Persian Dishes

Mohammad H. Eslami, Arshia Izadyari, Sadegh Mohammadian, Fateme Asgari, Ali Rahimi Akbar, Mohammad M. Vahedi

TOC

Introduction

Dataset

Model

Experiment

Evaluation

Challenges

Conclusion/Future Works

Introduction

- Retrieval → grounded answers in knowledge-intensive tasks
- Culinary domain: recipes, diet/ingredient queries, regional variants, image–text bridging
- Cross-modal food retrieval: Recipe1M, CLIP, M-CLIP
- RAG → retrieval + generation, reduces hallucinations
- Gaps: English-only focus, finished-dish photos, no Persian pipelines, label collisions
- Our work: Persian food-focused RAG with dual-modality retriever + generator
- Contributions
  - Ingredient-aware cross-model index
  - Label-aware contrastive training
  - Generator interface for Persian QA
  - Reproducible pipeline

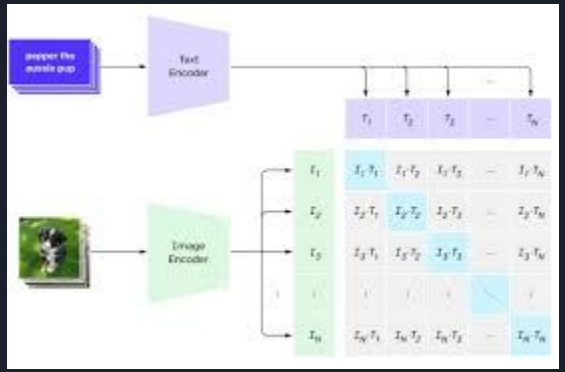
Dataset

- Web scraping using tools such as Selenium and BeautifulSoup
- Filled missing fields using LLMs
- Language and structure inconsistency
- Created documents and questions using LLMs
- Image as another modality
- Improved dataset by adding images of ingredients
- Image preprocessing done by model, only normalized the texts


```
{
  "title": "باقلوای اردبیل",
  "response": "ین کنید و از طعم بی‌نظیر باقلوا لذت ببرید
  "folder_path": "./final_foods/باقلوای_اردبیل"
},
```

Model (1)

- Shared Representation Learning
  - Vision Encoder
  - Text Encoder
  - Dimensional compatibility
- Retrieval and Evidence Pooling
  - Indexing
  - Initial retrieval
  - Option-aware expansion for MCQ



Model (2)

- RAG Prompt Construction
  - Snippet selection
  - Prompting modes
- Ingredient Photos
  - Training-time role
  - Inference-time behavior
- Answer Selection
  - Vision Language head
  - Heuristic head and fusion

Experiment(1)

- 50 MCQ questions
- 30 text-only questions
  - Food recommendation based on ingredients or location
  - Same foods from different locations
  - Meal type classification
  - Preparation instruction
  - Ingredients
- 20 text+image questions
  - Food classification
  - Geography of the food
  - Food recommendation
  - Identifying ingredients


```
{
    "question": "اسم این غذا چیه؟",
    "answer": "B",
    "image": "./2.jpg",
    "A": "کباب ترکی",
    "B": "آش دنگو",
    "C": "میگو سوخاری",
    "D": "خورشت به"
},
```



```
{
    "question": "خورش خلال کرمانشاهی با چه نوع خلالی شناخته می‌شود؟",
    "answer": "D",
    "A": "خلال سیب‌زمینی",
    "B": "خلال پسته",
    "C": "خلال هویج",
    "D": "خلال بادام"
},
```

Experiment(2)
Text-Only questions

Question:

درحال پخت حلوای نثار ارومیه هستم در چه مرحله‌ای حلوا را با دست ورز بدهم؟

Options:

بعد از سرد شدن کامل
بلافاصله بعد از اضافه کردن شکر
وقتی مواد کاملا داغ هستند
وقتی کمی ولرم شده

Answer:

وقتی کمی ولرم شده

Question:

غذا ی خوزستانی که با گیشنیز گوشی ماهی و تمر هندی درست می شود چیست؟

Options:

کوفته جنوبی
قلیه ماهی
فسنجان
میرزاقاسمی

Answer:

قلیه ماهی

Experiment(3)
Text+image questions

Question:

غذای داخل این عکس چیه؟

Options:

آش رشته
قطاب
قابلی پلو
قلیه ماهی

Answer:

قابلی پلو

Question:

مواد لازم این کباب چیست؟

Options:

بادمجان
سیب زمینی
گوشت و مرغ
ماهی

Answer:

گوشت و مرغ

Evaluation (1)

- Result Evaluation
  - Perform better when answering text-only questions
  - Limited number of images in dataset
  - Images are more difficult to encode
  - Harder to encode two modalities into the same space
- Models
  - Gemini 2.5 vs LLaVA 1.5
  - Retriever helps LLaVA 1.5
    - Less internal knowledge
  - Retriever works as an adversary for Gemini 2.5
    - Provides wrong knowledge

Generative Model	Model	Text-only	Image Question
LLaVA 1.5	Generative-only	26.67	20
	RAG	36.67	25
Gemini 2.5	Generative-only	53.33	40
	RAG	46.67	30

Evaluation (2)

- Question Type Evaluation
  - Text-only questions
    - Less reasoning required -> better performance
    - More keywords -> better performance

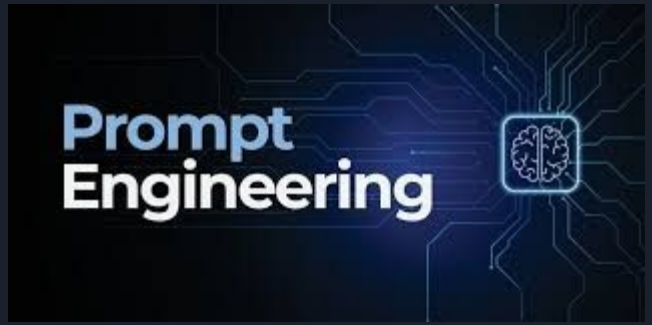
Generative Model	Model	recommendation	regions	type	prepartion	ingredients
LLaVA 1.5	RAG	2/9	0/2	2/2	2/5	5/12

  - Image + Text questions
    - Final food image as input -> better performance
    - Ingredient images are shared between foods
    - Simpler questions -> better performance

Generative Model	Model	classifying	geography	recommedation	ingredients
LLaVA 1.5	RAG	2/6	0/1	1/4	2/9

Challenges

- Data gathering
  - Limited fine-tuning data
  - Relying on APIs -> low quality of images and uncertainty about their correctness
- Data preparation
  - Choosing the appropriate preprocessing strategy
- Model development
  - Module selection
  - A small change in the model, big alternation in the results
  - Prompt engineering



Future Works

- Expanding the dataset with high-quality images
- Incorporating region-specific variations of dishes
- Trying more advanced retrievers and multimodal generators
- Integrating human-in-the-loop evaluation and feedback mechanism