A Dynamic Agent Framework for Large Language Model Reasoning for Medical and Visual Question Answering

Supplementary Material

6. 1. Neural Network in the Planning Module

The planning module adopts a 2-layer shallow neural network to realize the non-linear alignment between semantic embedding and agentic embedding. The architecture of neural network is as follows.

Table 4. Architecture of the Neural Network

Layer Type	**Description**
Input	Dimension = d_{in}
Hidden Layer(s)	Repeated for each $h_i \in$ `hidden_dims` Linear(d_{i-1}, h_i) BatchNorm1d(h_i) ReLU() Dropout(p)
Output	Linear(h_n, d_{out})

Table 5. Hyperparameters of the Neural Network

Hyperparameter	**Description**	**Value**
h_{in}	Input dimension	384
h_i	Hidden layer, n	2
h_{out}	Output dimension	2
`hidden_dims`	Hidden layer dimension	128, 56
`dropout`	Dropout rate	0.1
`learning_rate`	Learning rate	0.001

7. 2. Statistics of Cost-Effectiveness Analysis

Table 6. **Cost-Effectiveness Statistics in dataset VQARAD[16]** Summary of average token used, and processing time per questions across agents.

Models	Input Tokens	Output Tokens	Total Tokens	Time (second)
Model: GPT-4o-mini[23]				
Direct	144.142	77.7195	221.861	4.4168
CoT	219.397	162.828	382.224	7.5984
Visual	290.877	283.575	574.453	15.354
IICoT	656.192	1127.47	1783.66	22.215
MMCoT	746.989	1226.62	1983.62	23.783
MedDAF	256.3372	235.279	491.616	13.187
Model: Qwen-VL-plus[31]				
Direct	117.808	82.0083	199.817	4.6293
CoT	1299.65	133.957	2631.61	56.270
Visual	239.333	133.650	372.983	7.6572
IICoT	667.306	1217.58	1894.89	23.475
MMCoT	719.508	1253.59	1973.09	24.862
MedDAF	307.7682	152.927	460.694	6.4918
Model: Gemini-2.5 flash [8]				
Direct	144.198	78.6250	267.905	3.8676
CoT	338.492	332.350	660.842	14.390
Visual	404.126	839.172	1243.23	21.902
IICoT	733.446	2782.80	3516.24	39.210
MMCoT	651.564	2715.65	3367.21	71.990
MedDAF	348.9013	706.7585	1055.659	18.887

8. 3. Post-Hoc Case Analysis

This supplementary analysis presents a post-hoc case study examining the distribution pattern of aligned embeddings and demonstrating their alignment with practical meanings. To enhance visualization, we reduce the dimensionality to two dimensions, and for simplicity, we use the terms "left" and "right" in the subsequent description.

Case 1. The aligned embeddings of the questions are positioned on the left side of the agent embeddings IICOT, MMCOT. This indicates a clear preference for these two agents.

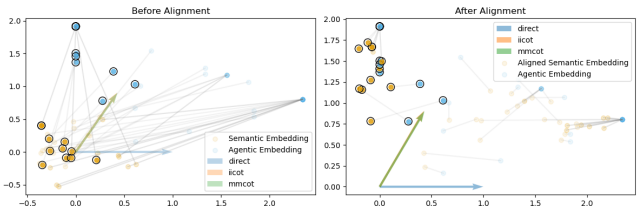

Figure 5. Case 1

Table 7. Sample questions in Case 1.

Index	Questions
1	Can you visualize the 4th ventricle?
2	The pathology seen in this image is typical of what disease?
3	Is there any blunting of the costophrenic angle(s)?
4	What is the vascular distribution of these infarcts?
5	Where is the pathology located?
6	Where are the infarcts located?
7	Can you see the abdominal fat pads?
8	Can a diagnosis or impression be made in this plane?

Case 2. The aligned embeddings are situated on the right side of the embedding space, suggesting a preference for direct inquiries without the inclusion of an image explanation (agent "direct").

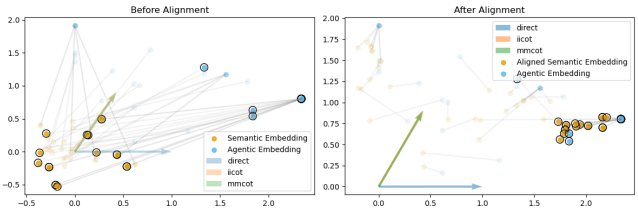

Figure 6. **Case 2.**

Table 8. Sample questions in Case 2.

Index	Questions
1	Is this confirmed appendicitis?
2	Does the skull appear fractured?
3	What disease does the pathology suggest?
4	Is there evidence of an abdominal hernia?
5	Is there evidence of herniation of the small bowel into the abdominal wall?
6	Where is the abnormality located?
7	Is this image sufficient to diagnose pericardial effusion?
8	Did the problem originate in the brain?

Case 3. The aligned embeddings are situated between the intermediate space among agents, illustrating the effectiveness from both two agents.

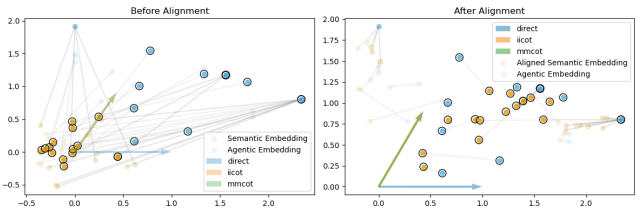

Figure 7. **Case 3.**

Table 9. Sample questions in Case 3.

Index	Questions
1	Is there a fracture of the left humerus?
2	What is the secondary abnormality in this image?
3	Is there a skull fracture?
4	Are there abnormal findings on this image?
5	In what parts of the brain are the lesions located?
6	Can you diagnose a pericardial effusion from this image?
7	Can you see the adrenals?
8	Are both lungs abnormally inflated?

Throughout the three cases, we identify the following observations:

Effectiveness of Self-Supervised Navigation on Agents. The cases show that learned question clusters align with the agent designs, under the condition that the learner (the planning module) is completely blind to each agent's settings. A clear difference is seen between questions that prefer visual-driven agents in case 1 (questions often start with "where", "is there", "can you see", . . .), and questions that prefer knowledge-only agents in case 2 (questions often start with "what", "is/does", "is there evidence", . . .). Although some questions are ambiguous to differentiate, the above patterns can suggest the efficacy of our models in matching suitable agents without relying on manual annotations.

Disentanglement from Questions' Semantic Embeddings to Agent-Performance Calibrated Embeddings. Our methods, being self-supervised, overcome a common shortcoming of routers, which is their over-dependence on the linguistic definition of agents. Posterior feedback terms play a crucial role in shaping the router. The subsequent cases demonstrate a distinct transformation from the initially sparsely distributed semantic embeddings to the converging agentic embedding. This demonstrates the effectiveness of feedback-based training in outperforming the forward-inference-only method.