AdGaze-3500: Evaluating Large Multimodal Models' Ability to Predict Human Attention to Ads

Appendix

This Appendix contains examples of the scanpaths generated by three LMMs (section 7), on the cleaning of that data before analysis (section 8.1), of the BERT-based prediction model (sections 8.2 and 8.3), and additional results from the experiments, including the full results and radar plots of the zero-shot predictions (section 9.1), and the dendrogram of the fine-tuned predictions (section 9.2).

7. Scanpaths generated by LMMs

7.1. Post-Processing

LMMs may produce flawed responses to the prompt specified in Section 4.1 of the main paper. There are three main types of mistakes and their corresponding corrections:

1. For very few models (e.g. Llama and Qwen), around 3-5 out of 3531 input images were mislabeled as NSFW (Not Safe For Work) contents, therefore requests for responses were declined. In this case, we filled the scanpath with an empty string and assigned 0 to all four gaze measures, which can be interpreted as LMMs deciding to skip the ad when reading.
2. In several cases, especially for open-sourced models (e.g. LLaVA family), the LMM could not stop generating and used up all output tokens. In these cases, we truncate the scanpaths to the last complete fixation description nearest to the end of the truncated output.
3. LMMs did not always perfectly conform to the required JSON format. In this case, we manually fixed the format by inspecting the generated response. All the issues were from misuses of quotes and commas; therefore, fixing them did not alter the semantic meanings.

7.2. Example Scanpaths

We show 3 example scanpaths generated by LMMs from 3 different families, including GPT 5, Qwen-VL plus, and MiniCPM-V 4.5. The sample scanpaths were generated with respect to a full-screen ad shown in Figure 5. The example scanpaths are shown in Table 4.

8. Fine-tuning Details

To accomplish Task 2 (Section 5.2 of the main paper), we trained a BERT-based regressor based on the scanpaths generated by LMMs. This section delineates all the steps for the fine-tuning task of interest.

Figure 5. The sample full-screen advertisement for which the sample scanpaths were generated.

8.1. Data Cleaning

We removed all predicted fixation durations from the generated scanpaths to eliminate any potential distractions and information leaking. The remaining contents remained in the JSON format for all LMMs tested.

8.2. BERT Model

We chose BERT-base-uncased [20] as the backbone interpreter of the scanpath strings. BERT-base-uncased is a bidirectional transformer-based encoder designed to learn deep bidirectional representations from unlabelled text. It contains 12 transformer blocks, hidden size of 768, and 12 attention heads. We retained the transformer encoder of BERT and discarded the Multi-Layer Perceptron (MLP) native to it.

8.3. MLP Regressor for Gaze Predictions

We replaced the BERT-native MLP by a trainable three-layer MLP regressor for predictions of the four gaze measures. In detail, our gaze prediction MLP consists of the following three layers:

1. `fc1 = nn.Linear(768, 384);`
2. `fc2 = nn.Linear(384, 384);`
3. `fc3 = nn.Linear(384, 4).`

The layers are interconnected by ReLU. Figure 1 in the main paper visualizes the architecture of our BERT-based gaze predictor.

Table 4. Three sample scanpaths generated for the Figure 5 by GPT-5, Qwen-VL plus and MiniCPM-V 4.5.

Model	Sample Scanpath

GPT-5	{
	"1": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Right page, woman's eyes and upper face with bright highlights against deep pink background",
	"Fixation Duration": 240
	},
	"2": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Right page, glossy lips and nose area with soft lighting",
	"Fixation Duration": 180
	},
	"3": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "Right page, white tagline 'YOU MAKE IT HAPPEN' thin sans-serif over dark pink",
	"Fixation Duration": 220
	},
	"4": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "Bottom-right, white serif logo 'LANCÔME PARIS'",
	"Fixation Duration": 200
	},
	"5": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Left page, metallic perfume bottle cap reflecting light",
	"Fixation Duration": 160
	},
	"6": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Left page, main body of pink perfume bottle with glow and reflections",
	"Fixation Duration": 220
	},
	"7": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "Left page, small word 'miracle' printed on the bottle",
	"Fixation Duration": 170
	},
	"8": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "Top-left headline 'miracle' in large white serif over pink sky",
	"Fixation Duration": 200
	},
	"9": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Left page, bright sun flare near bottle edge and horizon clouds",
	"Fixation Duration": 140
	},
	"10": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "Left page, white serif logo 'LANCÔME PARIS' near bottom of bottle",
	"Fixation Duration": 220
	},
	"Overall Gaze Duration (ms)": 1950
	}

Qwen-VL plus	{
	"1": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "Large white serif font spelling 'miracle' against a pink sky background",
	"Fixation Duration": 200
	},
	"2": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Clear perfume bottle with pink liquid, metallic cap, and reflective surface against a sunset sky",
	"Fixation Duration": 300
	},
	"3": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "Small black 'LANCÔME PARIS' text on the base of the perfume bottle",
	"Fixation Duration": 150
	},
	"4": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Woman's face with light skin tone, blue eyes, and blonde hair being touched by her hand",
	"Fixation Duration": 250
	},
	"5": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "White uppercase sans-serif font spelling 'YOU MAKE IT HAPPEN' across the lower right section",
	"Fixation Duration": 200
	},
	"6": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "White 'LANCÔME PARIS' text on black background in the bottom right corner",
	"Fixation Duration": 150
	},
	"7": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Pink and purple gradient sky with scattered clouds in the background of the left panel",
	"Fixation Duration": 100
	},
	"8": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "Hand of the woman touching her hair, showing fingers and skin texture",
	"Fixation Duration": 180
	}
	}

MiniCPM-V 4.5	{
	"1": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "Large white 'miracle' text at the top left of the ad",
	"Fixation Duration": 250
	},
	"2": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "The perfume bottle with a pink liquid and silver cap, centered on the left side",
	"Fixation Duration": 300
	},
	"3": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "LANCÔME PARIS logo below the perfume bottle",
	"Fixation Duration": 150
	},
	"4": {
	"Fixation Region Category": "picture",
	"Fixation Region Description": "The woman's face in the right half of the ad, her eyes and lips",
	"Fixation Duration": 350
	},
	"5": {
	"Fixation Region Category": "text",
	"Fixation Region Description": "YOU MAKE IT HAPPEN slogan near the bottom right of the image",
	"Fixation Duration": 200
	},
	"6": {
	"Fixation Region Category": "brand",
	"Fixation Region Description": "LANCÔME PARIS logo at the very bottom right corner",
	"Fixation Duration": 100
	}
	}

Table 5. Zero-shot prediction results of 20 LMMs on the four gaze times for all images in AdGaze-3500. In each column, the best score is double-underlined and the second best is underlined. The last row contains the average scores across all LMMs.

Gaze Type	Overall Ad			Brand Gaze			Text Gaze			Pictorial Gaze			Gaze Distribution	
LMM	*RMSE*	*RSE*	*Corr*	*RMSE*	*RSE*	*Corr*	*RMSE*	*RSE*	*Corr*	*RMSE*	*RSE*	*Corr*	*L_2 (std)*	*F-val (p-val)*
GPT 5 (high-thinking)	0.954	1.022	0.115	0.342	0.996	0.304	0.725	1.016	0.366	0.692	1.111	0.353	1.169 (.815)	4e2(.000)
GPT 5 (low-thinking)	0.987	1.057	0.110	0.334	0.974	0.332	0.754	1.057	0.348	0.708	1.137	0.359	1.192 (.858)	4e2(.000)
GPT mini	1.116	1.196	0.052	0.365	1.061	0.215	0.871	1.220	0.375	0.694	1.114	0.322	1.299 (.965)	4e2(.000)
Gemini 3 flash (thinking)	0.949	1.017	0.205	0.366	1.066	0.127	0.690	0.967	0.369	0.740	1.188	0.343	1.184 (.810)	4e2(.000)
Gemini 3 flash (non-thinking)	0.956	1.024	0.179	0.357	1.040	0.188	0.691	0.969	0.364	0.742	1.191	0.329	1.182 (.820)	4e2(.000)
Claude Sonnet 4.5 (thinking)	0.974	1.043	0.161	0.387	1.126	0.066	0.705	0.988	0.411	0.683	1.098	0.353	1.160 (.846)	4e2(.000)
Claude Sonnet 4.5 (non-thinking)	0.934	1.001	0.158	0.394	1.148	0.013	0.667	0.935	0.420	0.653	1.048	0.365	1.149 (.762)	4e2(.000)
Mistral 3 medium	1.001	1.072	0.196	0.372	1.082	0.157	0.693	0.971	0.425	0.645	1.036	0.314	1.259 (.671)	4e2(.000)
Mistral 3 small	1.001	1.073	0.166	0.364	1.059	0.297	0.695	0.974	0.403	0.720	1.156	0.337	1.202 (.831)	4e2(.000)
Llama 4 Maverick instruct	1.145	1.226	0.175	0.440	1.280	-0.052	0.732	1.026	0.413	0.794	1.276	0.308	1.321 (.962)	4e2(.000)
Llama 4 Scout instruct	1.785	1.912	0.118	0.465	1.353	0.174	1.021	1.431	0.293	1.065	1.711	0.231	2.077 (1.125)	4e2(.000)
Qwen 3 VL plus	1.149	1.231	0.084	0.363	1.056	0.288	0.786	1.101	0.337	0.795	1.277	0.288	1.323 (.976)	4e2(.000)
Qwen 3 VL plus (non-thinking)	1.161	1.244	0.152	0.396	1.154	0.146	0.720	1.009	0.353	0.766	1.231	0.298	1.347 (.893)	4e2(.000)
Qwen 3 VL flash	1.402	1.502	0.160	0.447	1.301	-0.043	0.885	1.241	0.313	0.895	1.438	0.238	1.621 (1.060)	4e2(.000)
MiniCPM-V 4.5	1.314	1.408	0.232	0.426	1.239	0.138	0.829	1.162	0.343	0.905	1.453	0.236	1.553 (1.002)	4e2(.000)
LLaVA vicuna v1.5 7b	1.993	2.134	0.041	0.892	2.597	0.037	1.180	1.653	0.081	0.985	1.582	0.086	2.092 (1.658)	4e2(.000)
LLaVA vicuna v1.6 7b	2.123	2.274	0.039	0.659	1.919	0.046	1.508	2.114	0.062	1.006	1.616	0.091	1.957 (2.097)	4e2(.000)
LLaVA vicuna v1.6 13b	3.691	3.954	0.005	0.808	2.353	0.018	2.496	3.498	0.017	1.792	2.877	0.110	2.426 (4.223)	4e2(.000)
Kimi k2.5 (thinking)	1.181	1.265	0.133	0.419	1.221	0.017	0.817	1.145	0.358	0.762	1.223	0.317	1.351 (.997)	4e2(.000)
Kimi k2.5 (non-thinking)	1.119	1.198	0.267	0.388	1.131	0.222	0.784	1.098	0.447	0.737	1.184	0.370	1.290 (.946)	4e2(.000)
Mean Performance	1.347	1.443	0.137	0.449	1.308	0.135	0.912	1.279	0.325	0.839	1.347	0.282	1.458 (1.166)	–

9. More Experiment Results

This section includes additional numerical results and plots for all experiments.

9.1. Task 1: Zero-shot Predictions

Full Results Table 5 shows the complete zero-shot results averaged over all 3,531 images. They lead to similar conclusions as those in the main paper regarding the zero-shot predictions:

1. Overall, LMMs do not show good capability in predicting gaze times, especially for ad gaze, which has the worst RMSE performance averaged over all 20 LMMs tested.
2. No single LMM achieved best predictions simultaneously for all four gaze measures. The L_2 metric again shows that Claude Sonnet 4.5 family models deliver the best overall predictions.

Radar Plots Figure 6 illustrates the 8 radar plots of different model families for the zero-shot predictions on the AdGaze-3500 test dataset. The plots support the previous conclusions from the Table 5, and also reveals that the zero-shot performances of models within each commercially-ready family, including all closed-source, Qwen and Kimi families, are almost identical. Among other families, the LLaVA family shows the largest variations in performance, with some of the models being unable to predict some of the gaze times at all. Surprisingly, LLaVA-vicuna-v1.6-13B, despite its larger scale and better training resources, performs worse than its predecessors. Further, we observe a larger deficiency of smaller models in predicting gaze times, compared to their sibling models of the same family, e.g. GPT-5-mini, Llama 4 Scout, and Qwen-VL flash.

Dendrogram Figure 3 in the main paper plots the dendrogram for the zero-shot LMM prediction results, based on the mean square error as the distance metric and using Ward's method [70] to link clusters. The dendogram shows further evidence that supports the conclusions dawn from Table 5 and Figure 6, as discussed in the main paper.

9.2. Task 2: Fine-tuned Predictions

Dendrogram. Figure 7 plots the dendrogram for the fine-tuned results based on the mean square error as the distance metric and using Ward linkage [70]. First, we observe that fine-tuning a BERT prediction model improves the prediction accuracy significantly for all 20 LMM models, as the range of the vertical axis, distance, decreased from 5 in the zero-shot setting (Figure 3) to 1.4: a 72% drop. At this smaller scale, the differences between the ground truth and LMMs' fine-tuned predictions are amplified, since the ground truth in the fine-tuning setting is the last to be clustered. Second, models from the same family, e.g. the GPT family, are generally clustered together; predictions from the closed-source models are similar to each other. At a higher level, all models, except the LLaVA models, present similar prediction capabilities. The LLaVA models, like in Figure 3, show up as outliers relative to the other models. These observations are consistent with the ones made in the Section 5.2 in the main paper.

Figure 6. Radar plots of the RMSE of the zero-shot predictions of four gaze measures for each of the 20 LMMs. Different models in the same family are shown in the same plot. The performance metrics are scaled to the $[0, 1]$ domain to enable comparisons, with higher values indicating better performance.

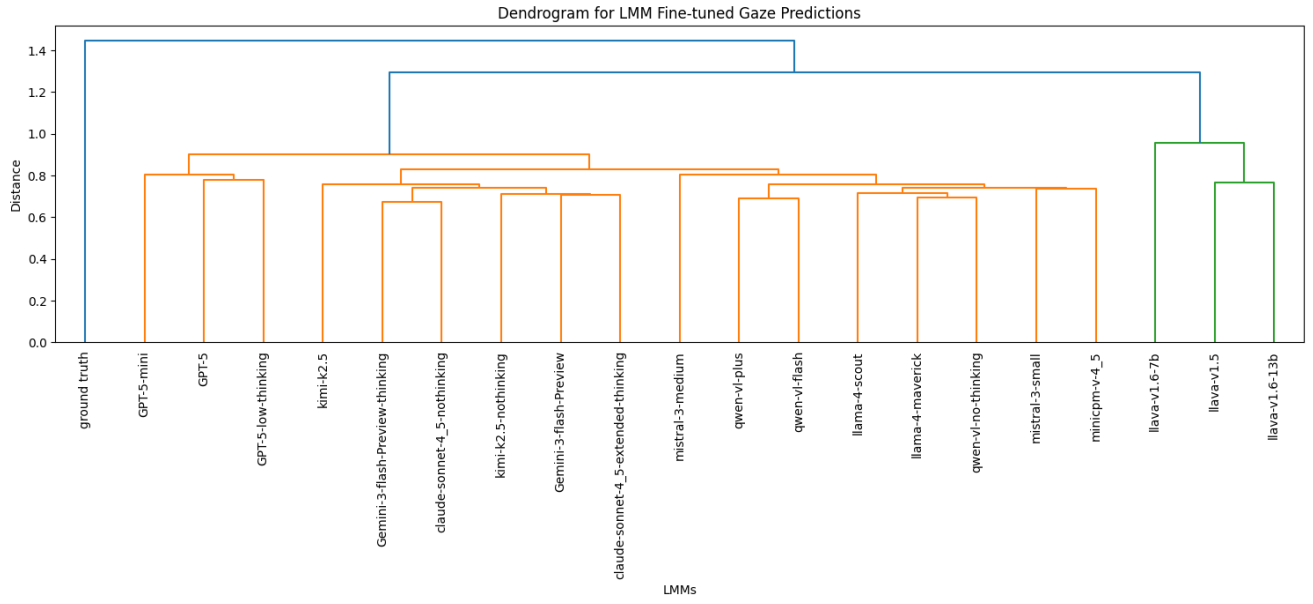

Figure 7. Dendrogram of the 20 LMM models and the ground truth based on fine-tuned predictions with mean square error as the distance metric and Ward linkage. The horizontal axis labels the LMMs and the ground truth. The vertical axis labels the distance.