1. AI Models

We evaluated four non-reasoning LLMs/MLLMs that can process both text and images. None of these models were fine-tuned for this specific task, but were directed prompted to perform the classifications using the same instructions given to human coders. Two of the models were employed via their APIs, while the other two used the Transformers Python library (Wolf, et al., 2020) via Hugging Face (https://huggingface.co/models?library=transformers).

1. GPT-4o (OpenAI) - A closed-source multimodal model capable of handling text and images in a single prompt. We accessed GPT-4o through the OpenAI API. It was selected for its strong general-purpose performance and ability to process ad images alongside captions and brand names. (https://openai.com/index/hello-gpt-4o/)
2. Pixtral-12B (Mistral AI) - A free-access VLM made available through the Mistral API. While smaller in scale than GPT-4o, it was included to assess how open-access, lower-cost models perform on the same classification task. (https://mistral.ai/news/pixtral-12b)
3. Qwen2.5-32B (Alibaba Cloud) - A high-capacity multimodal model run locally via the Transformers library on Maastricht University's Data Science Research Infrastructure GPUs. Its large size and open-weight availability make it an important benchmark for reproducible academic research. (https://huggingface.co/Qwen/Qwen2.5-32B-Instruct)
4. Gemma-3-12B (Google DeepMind) - Another open-weight model run locally on the same infrastructure. While smaller than Qwen2.5, it represents a newer generation of Google's Gemma family, designed for efficient inference without sacrificing multimodal capabilities. (https://huggingface.co/google/gemma-3-12b-it)

All four models were accessed in their base form without additional training or reasoning-mode activation. This ensures that performance differences reflect the models' out-of-the-box capabilities rather than task-specific adaptation. By including both commercial (GPT-4o) and open-weight (Pixtral, Qwen, Gemma) options, we were able to compare state-of-the-art proprietary models with openly available alternatives in the same classification setting.

1.1 AI Classification Instructions

You will be provided with a picture of potential outdoor advertising in Belgium. You will be given sets of questions about various aspects of the advertisement along with definitions and examples. Please answer each question using the exact format: `*QUESTION LABEL*: Yes/No – Brief explanation`*. Do not include any extra text, greetings, or commentary.*

For example, your answers should look like: `*CARTOON*: Yes/No – explanation`*;* `*CELEBRITY*: Yes/No – explanation`*; and so on. Ensure that the question label is between a set of stars. Ensure that each answer includes a brief*

explanation of the features in the image that led to your choice. Ensure that you answer all questions.

You will also be provided with a set of definitions which you should refer to when answering:

1. Food/drink manufacturing company or brand - a company or brand involved in producing and processing foods or beverages. Manufacturers focus on creating and packaging of consumable goods rather than selling directly to consumers. This category excludes restaurant/takeaway/delivery companies or brands and food retailers.

2. Food/drink retailer - a company or brand that sells food and drink products directly to consumers for home consumption (e.g., supermarkets, grocery stores, convenience stores and specialty food shops). These retailers primarily serve as intermediaries, providing a variety of products from different manufacturers to the end consumer. This category excludes manufacturing companies and restaurant/takeaway/delivery companies or brands.

3. Restaurant/takeaway/delivery outlet - a food service establishment that prepares and sells ready-to-eat meals and beverages for immediate consumption (either on the premises, through takeaway, or via delivery). These outlets focus on providing prepared food directly to customers for immediate or near-immediate consumption and do not primarily sell food or drink items in raw or packaged form for home cooking. This category excludes food retailers and manufacturing companies or brands.

4. Unprocessed or minimally processed food - natural foods (excluding Alcohol) that have undergone minimal changes (such as cleaning, drying, or freezing) without significant alteration to their nutritional content, e.g., fresh meat, eggs, frozen fruit.

5. Processed food - foods (excluding Alcohol) that have undergone processes like canning, smoking, fermentation or preservation, often with added ingredients to extend shelf life or enhance flavour, e.g., canned tomatoes, cheese, bread, smoked meat, dry fish.

6. Ultra-processed food - formulations of industrial ingredients (excluding Alcohol), resulting from a series of industrial processes such as frying, chemical modifications or application of additives, containing little or no whole foods, e.g., chips, candy, instant noodles, soft drinks, fast-food.

7. Processed culinary ingredients - substances (excluding Alcohol) extracted or refined from minimally processed foods, typically used in cooking or seasoning other food, e.g., sugar, butter, oils, spices.

List all distinct food items/dishes (not ingredients) observable in the image based on the overall presentation. Important: When the image features a composite food item (e.g., a burger, sandwich, or pizza), select only the single food category that best represents the

whole item rather than listing separate categories for its individual ingredients (such as bread, meat, or sauces).

For each [of the 23] food category, answer in the following format: `*QUESTION LABEL*: Yes/No – Brief explanation`. If you answer 'Yes' for a category, immediately provide a follow-up line indicating the level of processing for that food item in the format: `*<CATEGORY LABEL> PROCESSING*: <Processing Level> – Brief explanation`. Do not include any additional commentary. For example: `*CHOCOLATE SUGAR*: Yes – explanation; *CHOCOLATE SUGAR PROCESSING*: ULTRA PROCESSED – explanation`.

2. Qualtrics Survey

Figure S1 - Example of a Meta ad.

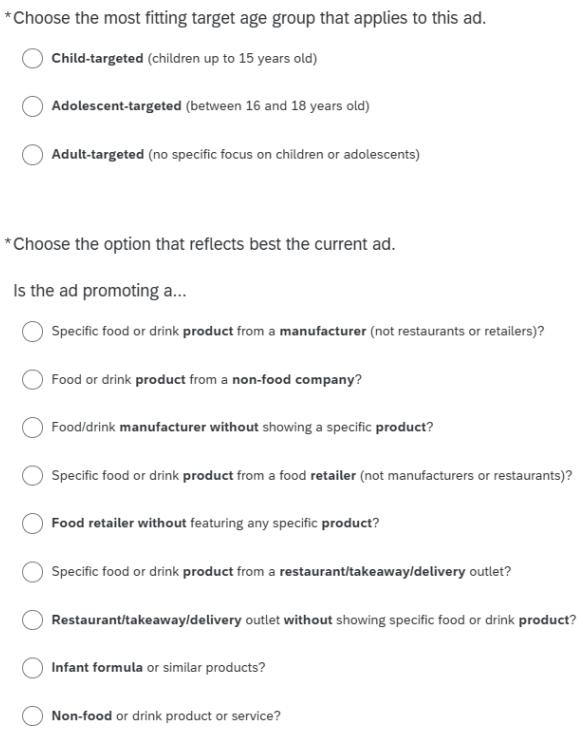

Figure S2 - Questions example from the Qualtrics survey.

3. AI Bias Analysis

This section presents the results of the AI Bias Analysis for the remaining single-option questions: Alcohol, Target Group and Ad Type.

Panel A (Alcohol) shows minimal bias across all models, with values close to zero. Panel B (Target Group) shows consistent significant under-detection of the *Adult* group across all models (ranging from -0.10 to -0.15), and an over-detection of *Child* content. The latter suggests that AI models are more likely than humans to identify ads as child-targeted. Panel C (Ad Type) shows more complex bias patterns, with most notable negative bias across all models for Manufacturer without a Product (from -0.06 to -0.15) and Restaurant without a

Product (from -0.08 to -0.10). Slight over-detection was observed for Restaurant with Product, but overall it seems the models struggled with identifying ads that contained no products.

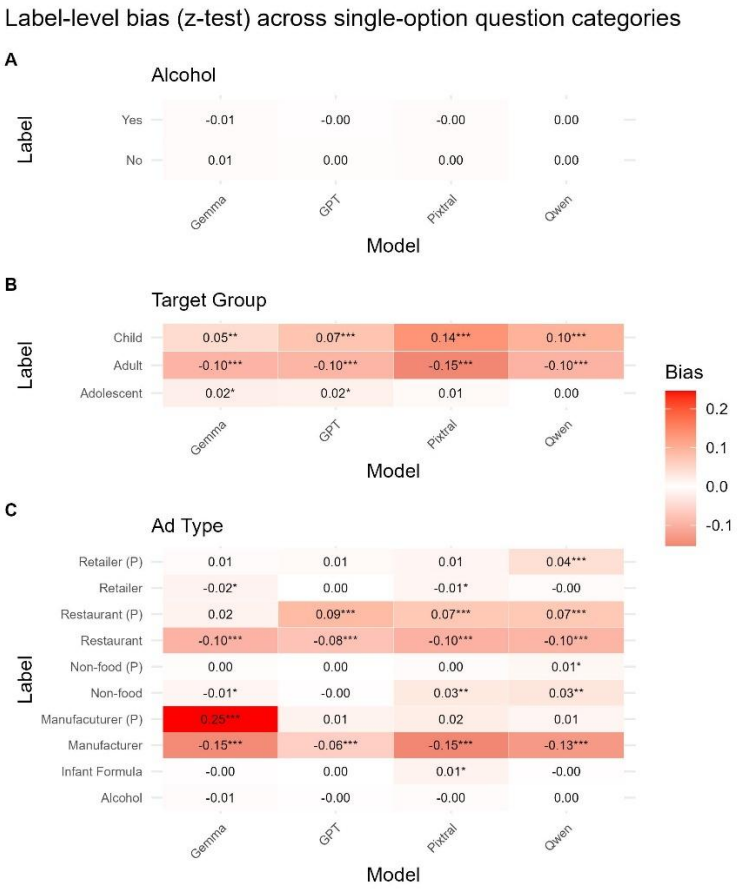

*Figure S3 - Label-level bias in AI model predictions for single-option questions compared to dietician consensus for **Panel A**: Alcohol, **Panel B**: Target Group and **Panel C**: Ad Type. The cells show option bias (mean selection rate difference: model - human) with z-test significance (*p < 0.05, **p < 0.01, ***p < 0.001). Positive values indicate over-selection; negative values indicate under-selection.*

4. Language Analysis

We compare agreement separately by language to examine whether certain classification tasks are more sensitive to linguistic context. To keep the analysis focused, we restrict attention to the two strongest-performing models, GPT and Qwen, to highlight the most relevant cross-language differences. The Figure below presents the delta in agreement for the four questions that showed most variation. The results show that for Marketing Strategies, both GPT and Qwen generally aligned more closely with dieticians and consensus in Dutch ads, with differences up to Δ ≈ +0.10. For WHO Categories, agreement was markedly higher in Dutch ads, especially for GPT and Qwen against the crowd consensus (Δ ≈ +0.12). Although the overall pattern points to Dutch ads being slightly easier for both GPT and Qwen to classify in multi-option settings, these differences were not statistically significant and insufficient to draw meaningful conclusions about language-specific variations.

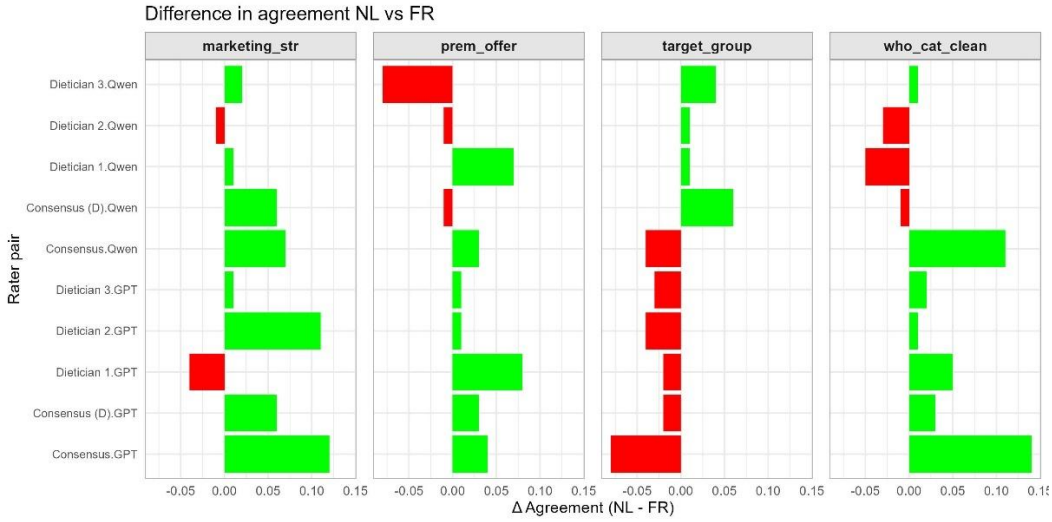

Figure S4 - Difference in agreement (Δ Krippendorff's Alpha for Marketing Strategies, Premium Offers and WHO Categories and Δ Gwet's AC1 for Target Group) between Dutch and French ads for GPT and Qwen compared to dieticians and consensus labels, across different tasks. Positive values indicate higher agreement in Dutch, negative values indicate higher agreement in French. None of the differences were statistically significant.

5. Outdoor Ads

Figure S5 – Example of an outdoor ad

For outdoor advertisements, GPT achieved consistently high agreement with dietician consensus across all single-option questions. Agreement was nearly perfect for Alcohol (0.97-0.99) and remained strong for Ad Type (0.69-0.80) and Target Group (0.85-0.96), the latter two showing an even higher agreement than for the Meta ads. These values were well

within the range of inter-dietician reliability, confirming that GPT performed comparably to human coders even in less-structured outdoor imagery.

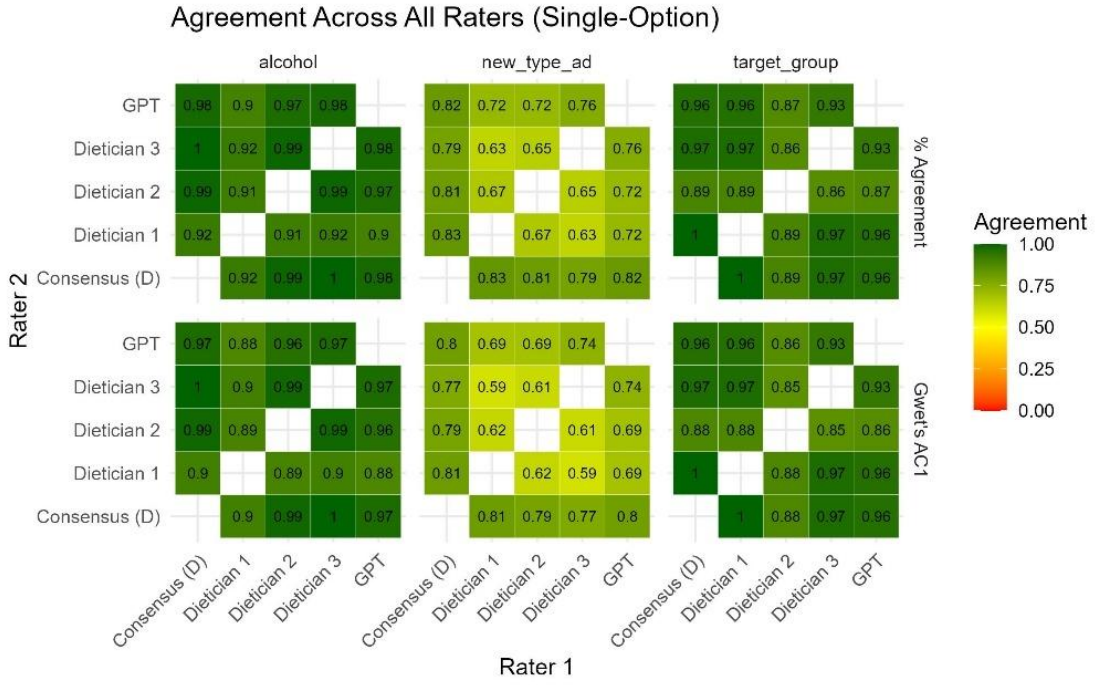

Figure S6 - Pairwise agreement between dieticians and GPT for single-option variables. Values represent both proportion agreement (top) and Gwet's AC1 (bottom). The agreement is computed over the full set of 100 outdoor ads.

As in the main dataset, agreement levels were lower for multi-option variables due to the higher complexity of co-occurring labels. Jaccard similarities typically ranged from 0.6 to 0.9 for Premium Offers and Marketing Strategies, and from 0.5 to 0.7 for WHO Food Categories, with corresponding Krippendorff's α values between 0.35 and 0.53. While this indicates moderate alignment, GPT's performance remained comparable to or slightly above inter-dietician agreement, suggesting robustness across both online and outdoor advertising contexts.

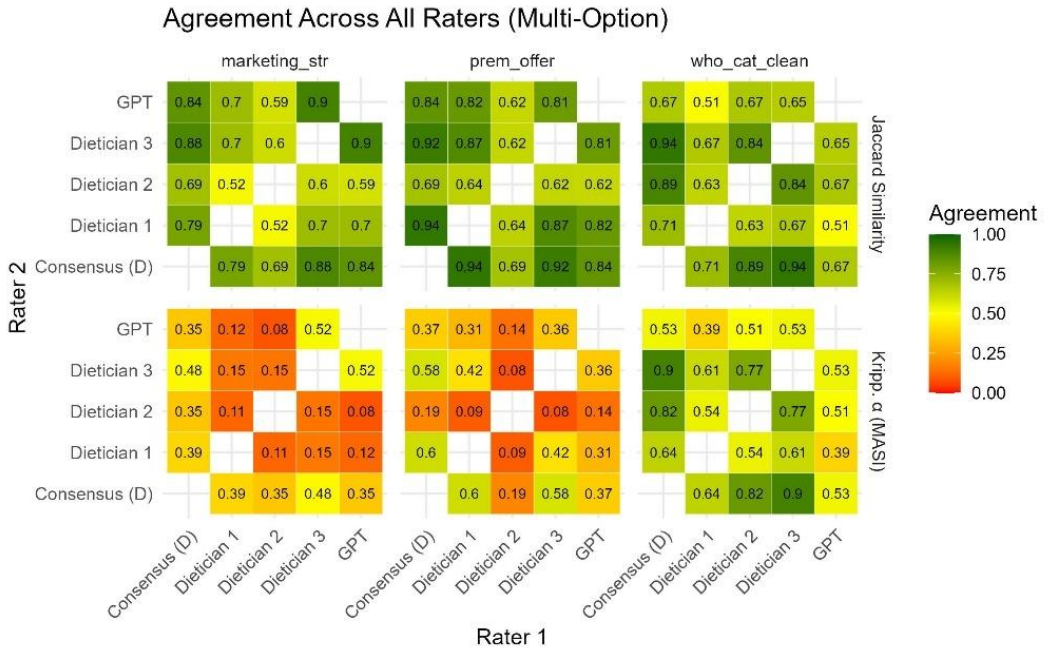

Figure S7 - Pairwise agreement between dieticians and GPT for multi-option variables. Values represent Jaccard similarity (top) and Krippendorff's Alpha with MASI distance (bottom). The agreement is computed over the full set of 100 outdoor ads.

For the additional Brand question included in the outdoor ad analysis, GPT reached a Jaccard similarity of 0.71-0.78 and a Krippendorff's α between 0.69-0.76 when compared to dietician consensus. These values closely matched inter-dietician agreement, indicating that GPT was similarly reliable in identifying brand presence or logos across diverse outdoor ad contexts.

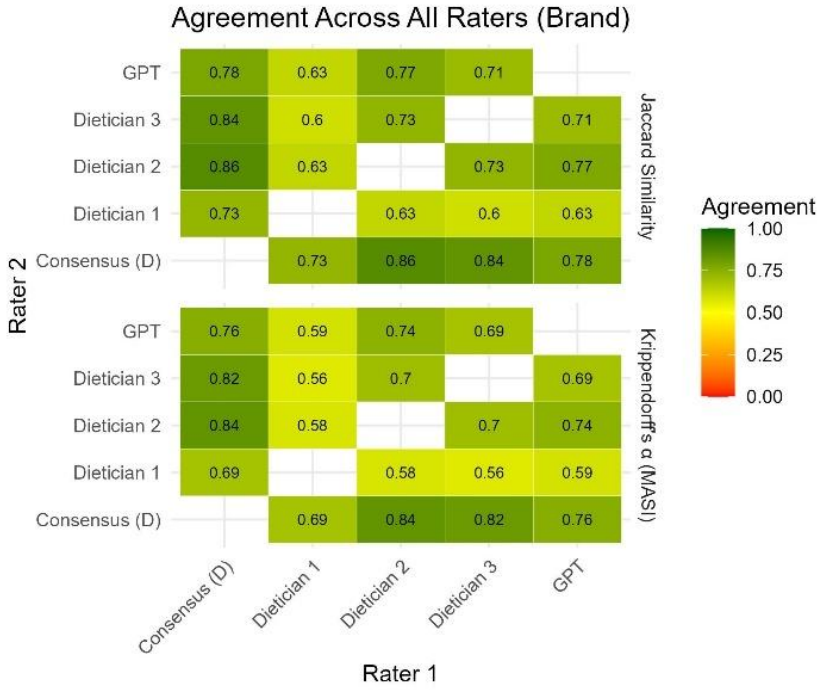

Figure S8 - Pairwise agreement between dieticians and GPT for the brand question. Values represent Jaccard similarity (top) and Krippendorff's Alpha with MASI distance (bottom). The agreement is computed over the full set of 100 outdoor ads.

References

Wolf, T. et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38-45. Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/2020.emnlp-demos.6