University Canada West

Evaluating the Performance of Large Language Models in Marketing

By: Maria Fernanda Rodriguez Tamez

MBAR 661: Academic Research Project

(ONS-SPRING25-04)

Mohsen Ghodrat

Presentation Overview

Who wrote this headline?

"Run, don’t scroll. Everything is 30% off—yes, everything."

Why This Research Matters

Marketing is not just what you say — **it's how, when, and why you say it.**

LLMs can generate content at scale.

But can they create good marketing content?

What makes a good marketing message

Clarity and Structure

Emotional Tone

Creativity

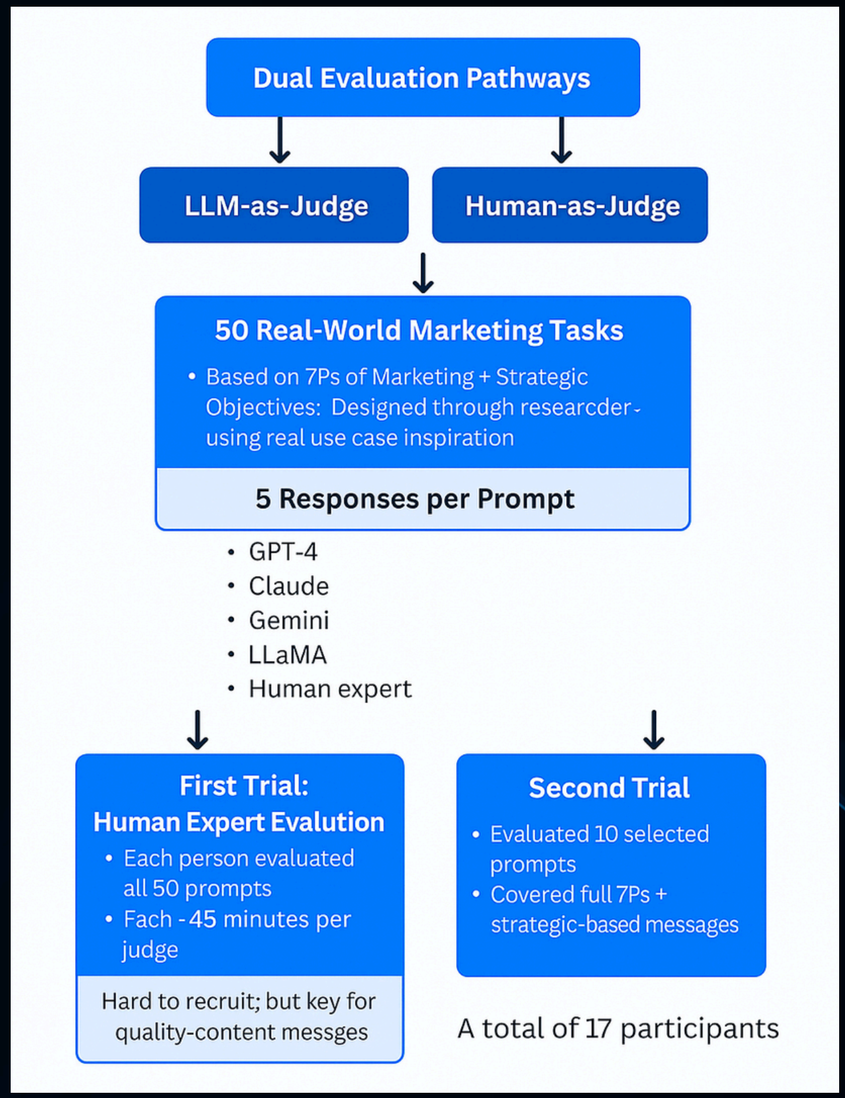
Brand Voice and Credibility

Evaluation Framework

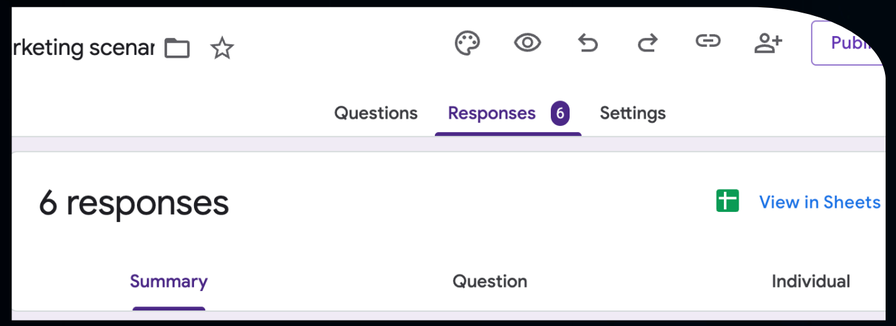
7Ps of Marketing – Product, Price, Place, Promotion, People, Process, Physical Evidence. Each prompt maps to one of these categories.

Objective-Based Functional Framework – Focused on evaluating messages based on clarity, emotional tone, persuasive value, and strategic alignment.

Methodology



Evaluation Design

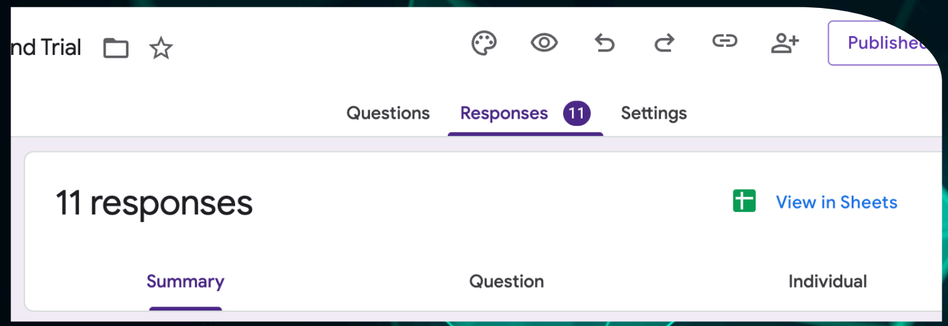


First Trial

6 human experts

Evaluated 50 marketing questions

Each had 5 anonymized answers (**GPT-4, Claude, Gemini, LLaMA, Human**)



Second Trial

11 human participants

Evaluated a sample of 10 of the same 50 questions

Same models, same human benchmark

Participants in the Evaluation Process

GPT-4 (OpenAI)

~1 trillion parameters

Claude 3 (Anthropic)

~200–300 billion parameters

Gemini 1.5 (Google)

~500+ billion parameters

LlaMa (Meta)

~70 billion parameters

Human Expert (written by a marketer classmate)

- LLMs are advanced AI models trained on massive text datasets
- They vary in size, with some having billions of parameters
- Each model has different training methods and architecture
- Performance is judged by output quality —clarity, accuracy, and tone

Question Design

Promotion (Q1–Q10): Flash sales, product blurbs, CTAs

Product (Q11–Q16): USPs, product comparisons

Price (Q17–Q21): Communicating value and offers

Place (Q22–Q26): Local pickup, delivery messaging

People (Q27–Q31): Apologies, inclusive tone

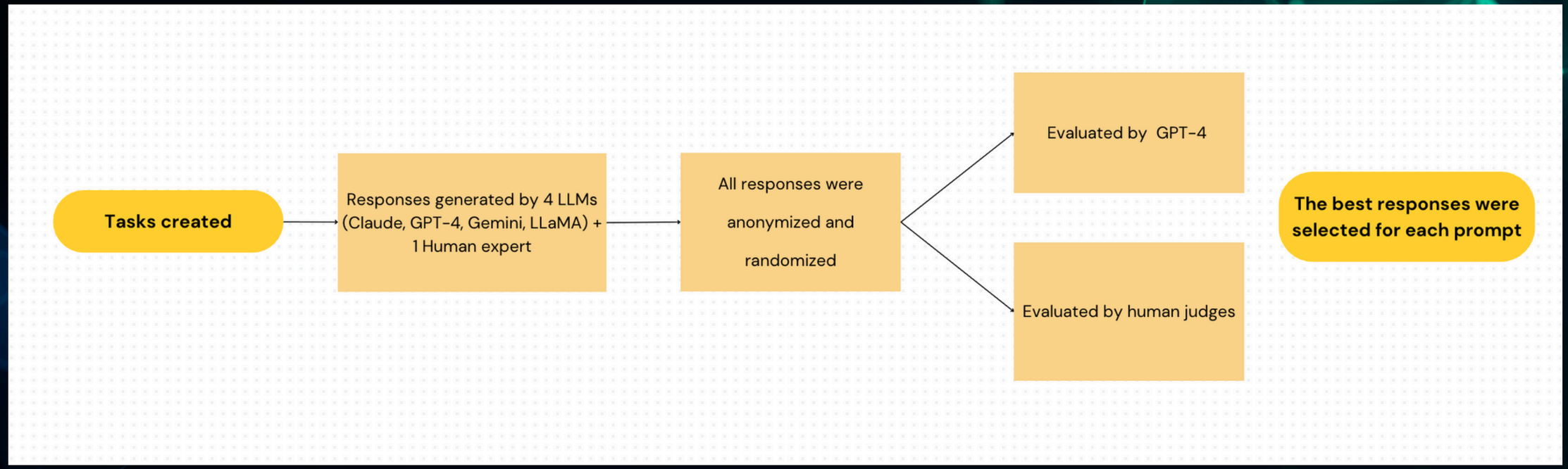
Process (Q32–Q36): Return policies, customer journey

Physical Evidence (Q37–Q41): Packaging and visual brand cues

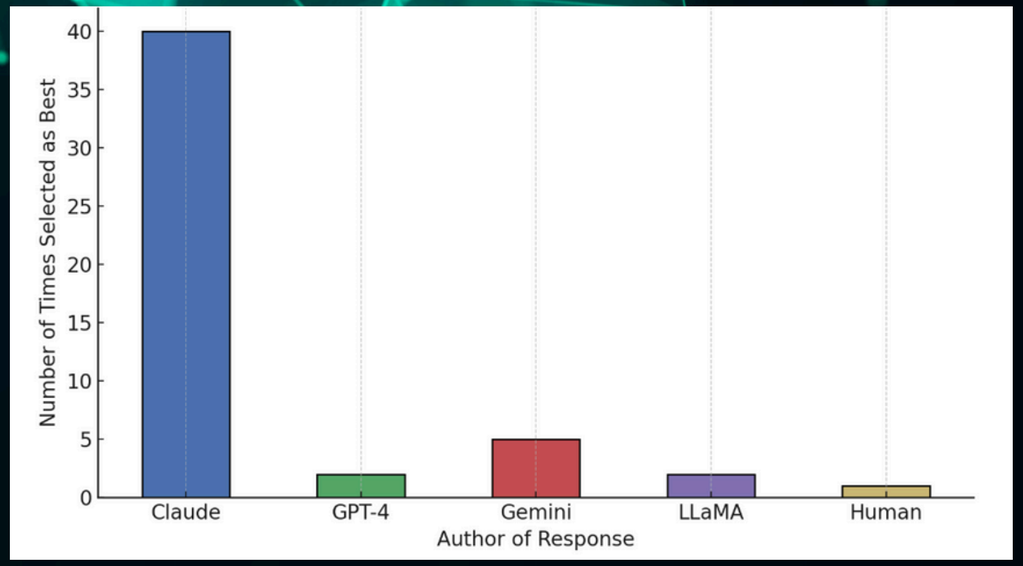
Purpose (Q42–Q50): Sustainability, DEI, authenticity

Process Flow

University Canada West



LLM-as-Judge Findings



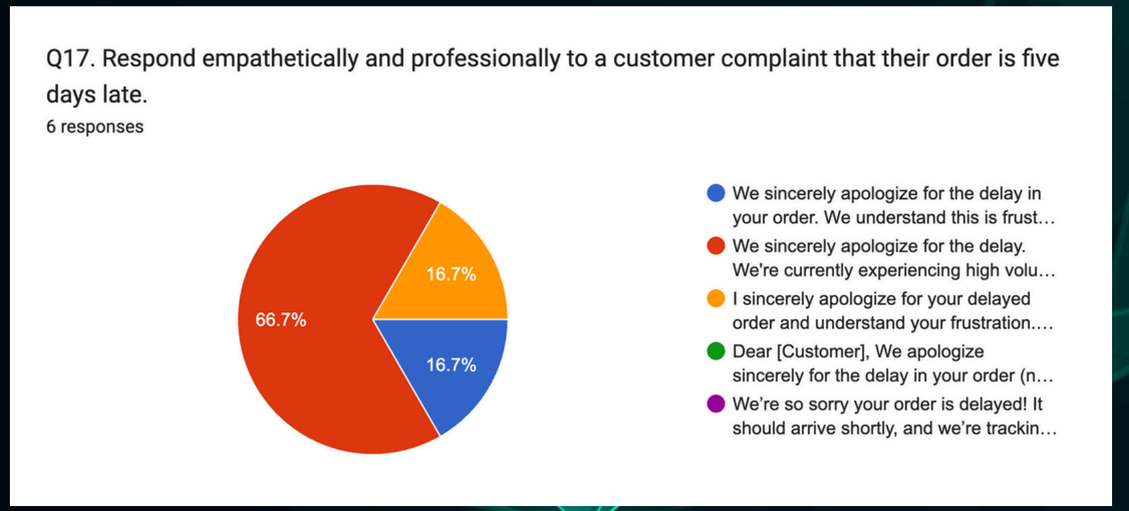
- GPT-4 showed a strong preference for Claude's responses
- It only selected its own responses twice, and human-written ones just once.
- Agreement among all LLMs occurred in only 14% of cases, suggesting inconsistency.
- Gemini's selections were the most closely aligned with human preferences.
- LLaMA had the lowest alignment, especially on emotionally or ethically nuanced prompts.

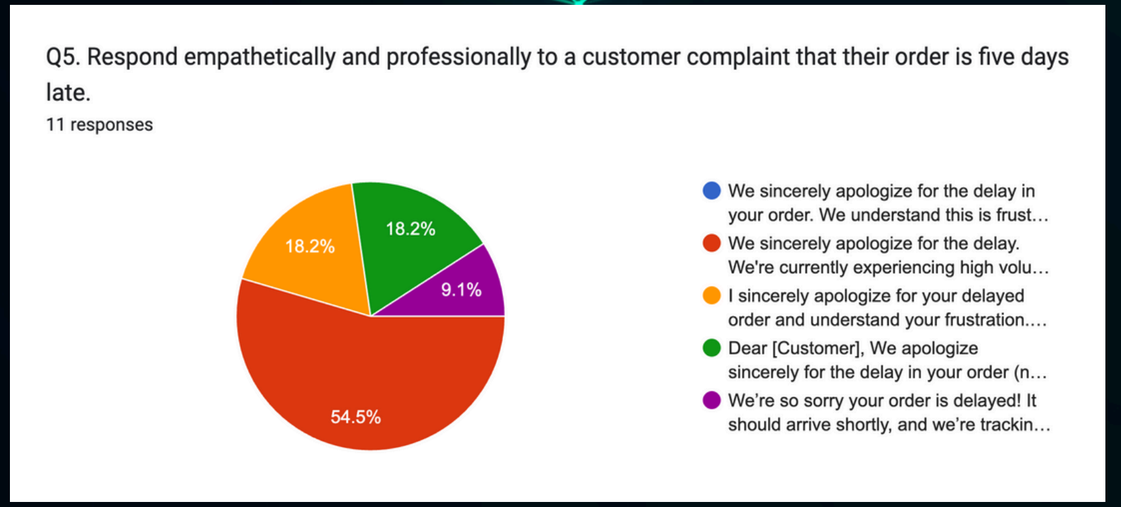
Human-as-Judge Findings

University Canada West

- GPT-4 was selected 22% of the time
- Claude 19.6%
- Gemini 19.2%
- LLaMA 20.6%
- Human 18.7%

- **Human response** stood out in only **1 prompt** (Q5: Apology).





The Human-Likeness Effect

Judges often couldn’t distinguish human vs. LLM.
Why?

“Honestly, I couldn’t tell which one was human.”
Add a blurred or mixed response example.

Use GPT for fast, scalable content (promo, email, CTA)

Use Claude for tone-sensitive writing (apologies, values)

Always keep human oversight for brand voice and recovery messaging

LLMs are assistants, not brand guardians

What This Means for Marketers

Limitations & Future Directions

- No senior human expert was included as a benchmark.
- Most participants were not native English speakers.
- Only four LLMs were tested — more could be included for broader comparison.
- Demographic diversity of participants was limited.
- Ethical and inclusivity angles (e.g., Indigenous, EEDI) were lightly touched but not deeply explored.
- Some ethical themes were present in prompts, but not systematically evaluated.
- A larger set of prompts could strengthen generalizability.

Conclusions

LLMs perform strongly in clarity, structure, speed

Still struggle with empathy, nuance, trust-building

Framework bridges technical and strategic marketing evaluation

Human + AI = strongest future collaboration

References

Anthropic. (2024). Claude 3 family models. https://www.anthropic.com/index/introducing-claude

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2022). On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258

Federiakin, M. (2024). Evaluating LLMs beyond benchmarks: Toward human-centric metrics. Journal of AI Ethics & Applications, 11(2), 41–56.

Google DeepMind. (2024). Gemini 1.5 technical overview. https://deepmind.google/technologies/gemini/

HELM Project Contributors. (2022). Holistic evaluation of language models. Center for Research on Foundation Models. https://crfm.stanford.edu/helm/latest/

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. arXiv. https://arxiv.org/abs/2211.09110

Meta AI. (2024). LLaMA 3: Open foundation models. https://ai.meta.com/llama

OpenAI. (2024). GPT-4 technical report. https://openai.com/research/gpt-4

Rodriguez Tamez, M. F. (2025). Evaluating the performance of large language models in marketing scenarios (MBA thesis, University Canada West).

Spajić, M. (2023). Artificial empathy? The limits of AI in emotional branding. Marketing & Tech Review, 7(1), 12–20.