Comparison of Large LLMs for Question Generation

Evaluation Criteria (0–5 Scale)

Each model is evaluated on:

- **Relevance**: Are the questions grounded in the passage content?
- **Diversity**: Are the 3 questions semantically distinct?
- **Clarity**: Is the language fluent and coherent?
- **Answerability**: Can the passage answer the question?
- **Instruction-Following**: Does it follow the "Generate 3 questions" format?

Model Comparison Table (0–5)

Model Name	Params	Type	Rel.	Div.	Clarity	Answer	Follows Inst.
GPT-4.5 / GPT-4-turbo	∼1.8T	Closed	5.0	5.0	5.0	5.0	5.0
Claude 3 Opus	∼1T	Closed	5.0	4.5	5.0	4.5	5.0
Command R+ (Cohere)	35B	API / HF	4.5	4.5	4.5	4.5	4.5
Meta-LLaMA-3-8B-Instruct	8B	Open	4.5	4.0	4.5	4.0	4.0
Mixtral-8x7B-Instruct	12.9B	Open	4.5	4.5	4.5	4.5	4.5
oh-dcft-v3.1-claude-3-5-haiku	8B	Open Clone	4.0	3.5	4.0	3.5	4.0
Mistral-7B-Instruct-v0.2	7B	Open	4.0	3.5	4.0	4.0	4.0
Gemma-7B-it	7B	Open	3.5	2.5	3.0	2.5	3.0

Focus: `oh-dcft-v3.1-claude-3-5-haiku`

- **Relevance:** 4.0 — Mostly grounded in passage
- **Diversity:** 3.5 — Can repeat concepts or phrasing
- **Clarity:** 4.0 — Fluent and readable
- **Answerability:** 3.5 — Sometimes vague
- **Instruction-following:** 4.0 — Usually adheres to format

Recommendations

- **Best Overall Quality:** GPT-4.5 or Claude 3 Opus
- **Best Free Open-Weight Model:** Meta-LLaMA-3-8B-Instruct
- **Most Efficient for Colab:** Mistral-7B-Instruct
- **Best Long-Passage Model:** Mixtral-8x7B-Instruct (requires large VRAM)
- **Best Claude-like Free Model:** `oh-dcft-v3.1-claude-3-5-haiku`

Conclusion

The `oh-dcft-v3.1-claude-3-5-haiku` model is a reasonable Claude-style free alternative with solid performance on question generation, though slightly behind top-tier models in reasoning and diversity.