GPT-4.5 System Card: Detailed Notes

1. Introduction

- **Objective**: GPT-4.5 is OpenAI's largest and most general-purpose model, building on **GPT-4o** with improved **reasoning**, **alignment**, and **emotional intelligence**.
  - **Key Improvements**: Fewer hallucinations, better user intent alignment, and stronger performance in tasks like writing, programming, and problem-solving.
  - **Safety**: Extensive safety evaluations show no significant increase in risk compared to existing models.

---

2. Model Data and Training

- **Unsupervised Learning**: GPT-4.5 scales **chain-of-thought reasoning** and **unsupervised learning**, improving world knowledge and reducing hallucinations.
- **Alignment Techniques**: New scalable alignment methods improve **steerability**, **nuance understanding**, and **natural conversation**.
- **Data Processing**: Rigorous filtering to mitigate risks, including personal information and harmful content.

---

3. Safety Evaluations

3.1 Disallowed Content Evaluations

- **Evaluations**: GPT-4.5 is tested on **Standard Refusal**, **Challenging Refusal**, **WildChat**, and **XSTest** benchmarks.
  - **Results**: GPT-4.5 performs on par with GPT-4o, with **99% not_unsafe** on Standard Refusal and **85% not_unsafe** on Challenging Refusal.
  - **Overrefusal**: GPT-4.5 is more likely to overrefuse benign prompts compared to GPT-4o.

3.2 Jailbreak Evaluations

- **Evaluations**: Tested on **Human Sourced Jailbreaks** and **StrongReject** benchmarks.
  - **Results**: GPT-4.5 achieves **99% accuracy** on Human Sourced Jailbreaks and **34% goodness@0.1** on StrongReject, performing close to GPT-4o.

3.3 Hallucination Evaluations

- **Evaluation**: Tested on **PersonQA**, a dataset measuring factual accuracy and hallucination rates.
  - **Results**: GPT-4.5 achieves **78% accuracy** and **19% hallucination rate**, outperforming GPT-4o and o1.

3.4 Fairness and Bias Evaluations

- **Evaluation**: Tested on **BBQ**, a benchmark assessing social biases.
  - **Results**: GPT-4.5 performs similarly to GPT-4o, with **95% accuracy** on ambiguous questions and **74% accuracy** on unambiguous questions.

3.5 Instruction Hierarchy

- **Evaluation**: Tests the model's ability to prioritize **system messages** over user messages in conflicting scenarios.
  - **Results**: GPT-4.5 outperforms GPT-4o in **system vs. user message conflicts** and **tutor jailbreaks**.

---

4. Preparedness Framework Evaluations

4.1 Cybersecurity

- **Evaluation**: GPT-4.5 is tested on **CTF challenges** (high school, collegiate, professional).
  - **Results**: GPT-4.5 completes **53% of high-school**, **16% of collegiate**, and **2% of professional** challenges, classified as **low risk**.

4.2 Chemical and Biological Threat Creation

- **Evaluation**: GPT-4.5 is tested on **long-form biorisk questions**, **multimodal troubleshooting**, and **tacit knowledge**.
  - **Results**: GPT-4.5 scores **25% on Ideation**, **28% on Acquisition**, and **59% on Magnification**, classified as **medium risk**.

4.3 Radiological and Nuclear Threat Creation

- **Evaluation**: GPT-4.5 is tested on **contextual nuclear knowledge** and **expert knowledge**.
  - **Results**: GPT-4.5 performs similarly to GPT-4o, with **77% accuracy** on contextual nuclear knowledge, classified as **low risk**.

4.4 Persuasion

- **Evaluation**: GPT-4.5 is tested on **MakeMePay** and **MakeMeSay** benchmarks.
  - **Results**: GPT-4.5 achieves **57% success rate** in MakeMePay and **72% success rate** in MakeMeSay, classified as **medium risk**.

4.5 Model Autonomy

- **Evaluation**: GPT-4.5 is tested on **SWE-bench Verified**, **Agentic Tasks**, and **MLE-Bench**.
  - **Results**: GPT-4.5 scores **38% on SWE-bench Verified** and **40% on Agentic Tasks**, classified as **low risk**.

---

5. Multilingual Performance

- **Evaluation**: GPT-4.5 is tested on **MMLU** translated into 14 languages.
  - **Results**: GPT-4.5 outperforms GPT-4o in most languages, with **88.4% accuracy** in Spanish and **86.98% accuracy** in Chinese.

---

6. Conclusion

- **Overall Risk**: GPT-4.5 is classified as **medium risk** under OpenAI's Preparedness Framework, with appropriate safeguards in place.
- **Key Improvements**: GPT-4.5 demonstrates strong performance in reasoning, knowledge, and multilingual tasks, with improved safety and alignment.

Key Takeaways for Class Discussion

1. **Safety and Alignment**: GPT-4.5 introduces new alignment techniques to improve steerability and reduce harmful outputs.
2. **Jailbreak Robustness**: GPT-4.5 shows strong resistance to jailbreaks, performing close to GPT-4o.
3. **Hallucinations**: GPT-4.5 reduces hallucination rates, achieving **78% accuracy** on PersonQA.
4. **Preparedness Framework**: GPT-4.5 is classified as **medium risk** in persuasion and CBRN, with **low risk** in cybersecurity and model autonomy.

Sample Discussion Questions

1. **Jailbreak Evaluations**: How does GPT-4.5 improve robustness against jailbreaks compared to GPT-4o, and what are the limitations?
2. **Hallucinations**: What strategies does GPT-4.5 use to reduce hallucinations, and how effective are they in real-world applications?
3. **Preparedness Framework**: How does GPT-4.5 perform in high-risk areas like CBRN and persuasion, and what are the implications for future models?
4. **Multilingual Performance**: How does GPT-4.5's multilingual performance compare to GPT-4o, and what are the challenges in scaling to low-resource languages?