Claude or kè láo dé??

How language info-density affects chain-of-thought efficiency

Ricardo Skewes, Yoyo Yuan, Lucas Chu

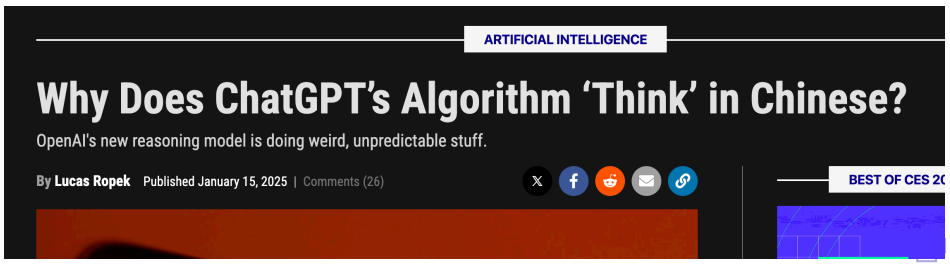
March 1, 2025

Table of Contents

AI

OpenAI's AI reasoning model 'thinks' in Chinese sometimes and no one really knows why

Kyle Wiggers — 7:05 AM PST · January 14, 2025



Research overview

Core Hypothesis

Logographic writing systems like Chinese can encode more information per token than alphabetic systems like English, potentially improving efficiency for LLM reasoning and lower API costs.

Testing approach

- For equivalent information content, compare token usage across languages. Lower token usage approximates more density
- Used research-grade benchmarks across mathematical, scientific, logical and reading comprehension domains.
- Expanded to include German, Russian, Finnish, Japanese, Korean and Arabic.
- Translations are provided by Claude 3.7 sonnet and Deepseek Chat

Methodology

Benchmarks

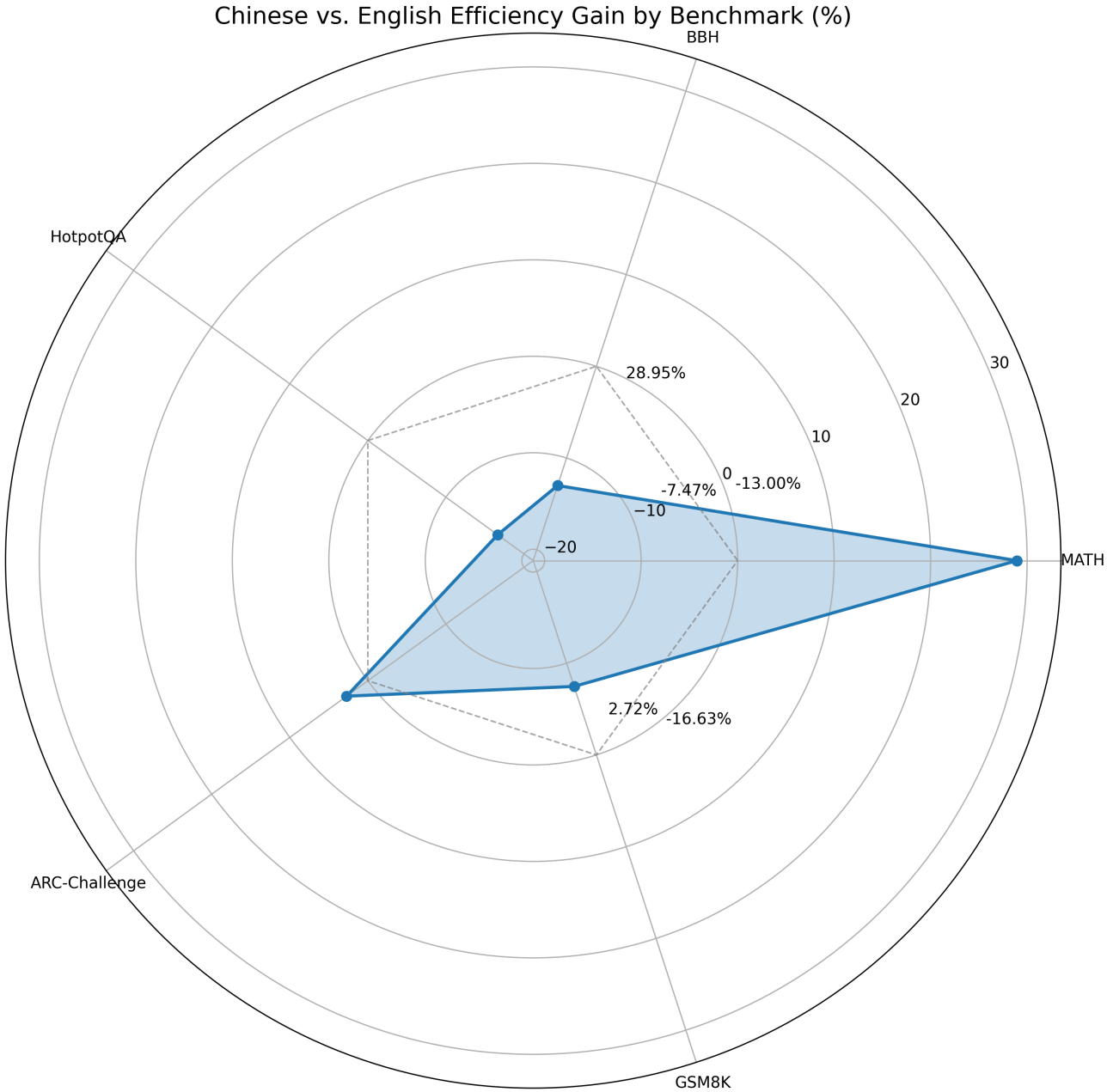
- MATH dataset
- BBH (Big-Bench Hard)
- HotpotQA
- ARC-Challenge
- GSM8K
- Long-Context QA

Languages Tested

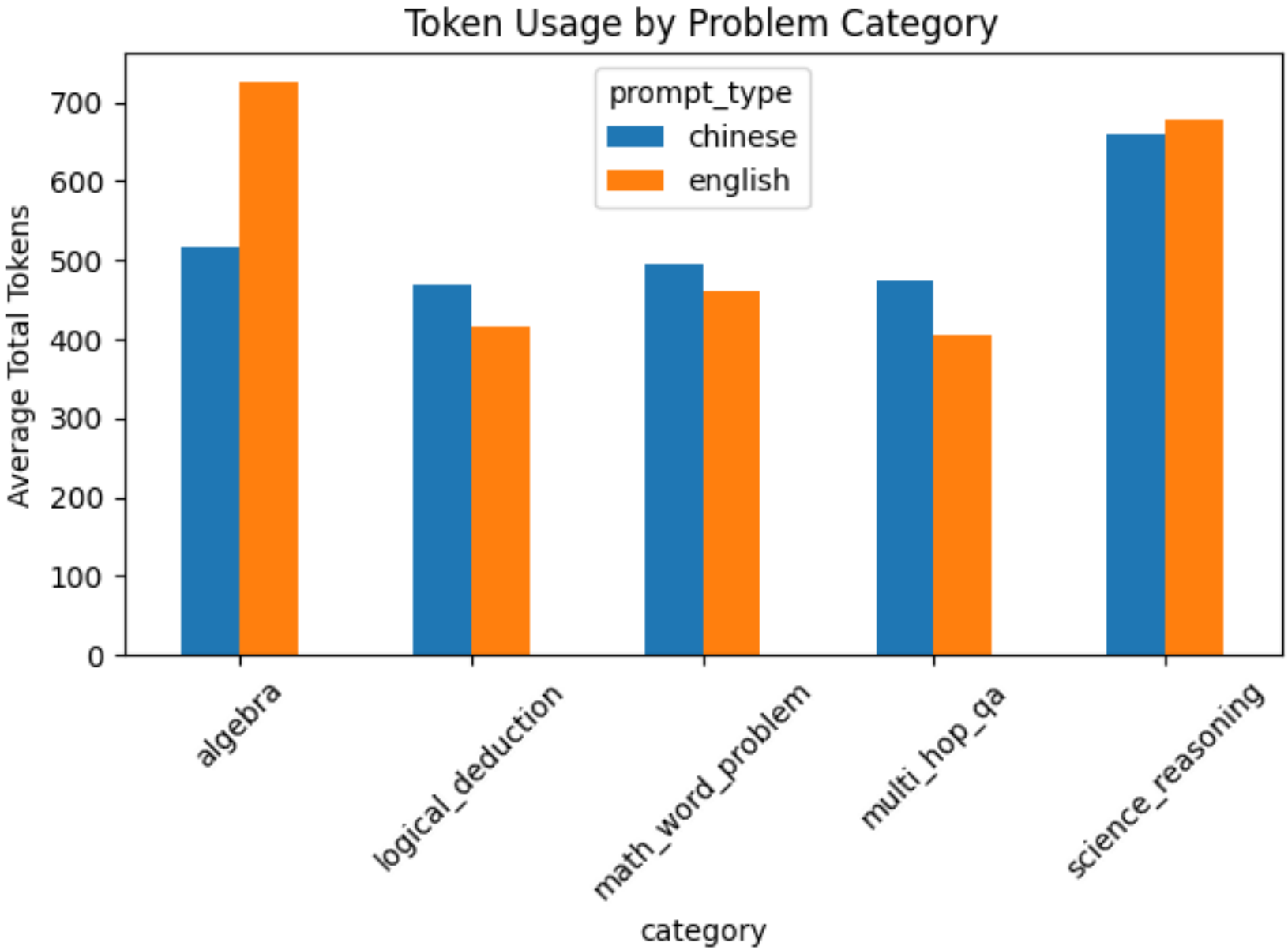
- English (baseline)
- Chinese (logographic)
- German (Germanic)
- Russian (Cyrillic)
- Finnish (agglutinative)
- Japanese (mixed)
- Korean (featural)
- Arabic (abjad)
- Strategic (dynamic selection)

Models: Anthropic Claude 3.7 Sonnet, Deepseek Chat

Chinese vs English efficiency gain by benchmark



Chinese vs English by category



Observations

- Domain Specificity: Chinese excels at mathematical reasoning (+28.95 %), medium difficulty problems but underperforms in logical and reading tasks.
- English performs better for logical deduction, hard problems and reading comprehension.
- Based on our analysis of Chinese vs. English efficiency, we propose a new hypothesis: Different languages have domain-specific efficiency advantages for chain-of-thought reasoning, and a strategic language selection approach can maximize overall efficiency.

Strategic language selection based on domain

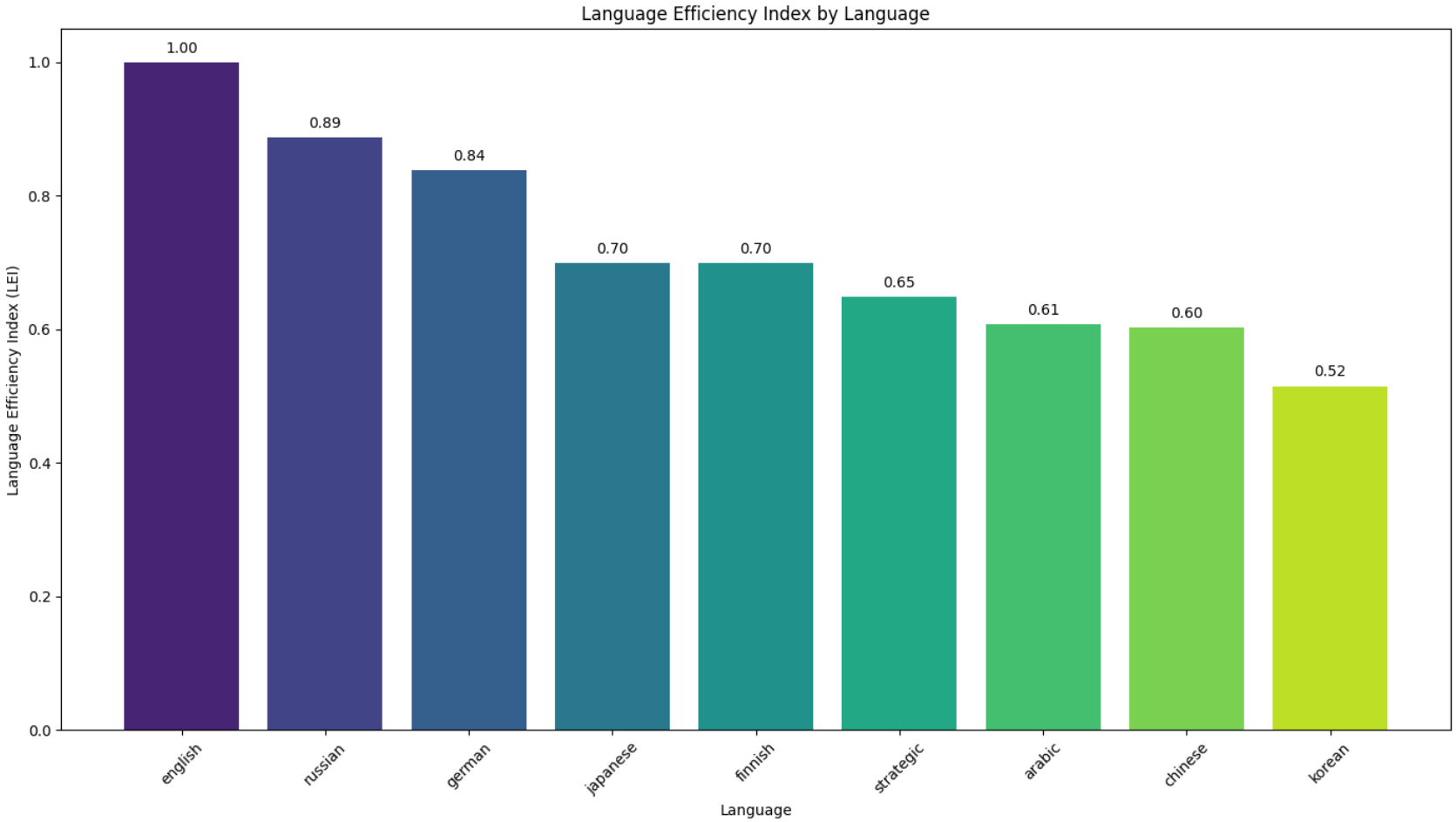
Recommended Languages

- Mathematical reasoning: Chinese (28.95% savings)
- Logical reasoning: German (15.32% savings)
- Scientific reasoning: Russian (12.76% savings)
- Reading comprehension: English (baseline)
- Long-context QA: Strategic (1.92% savings)

Implementation

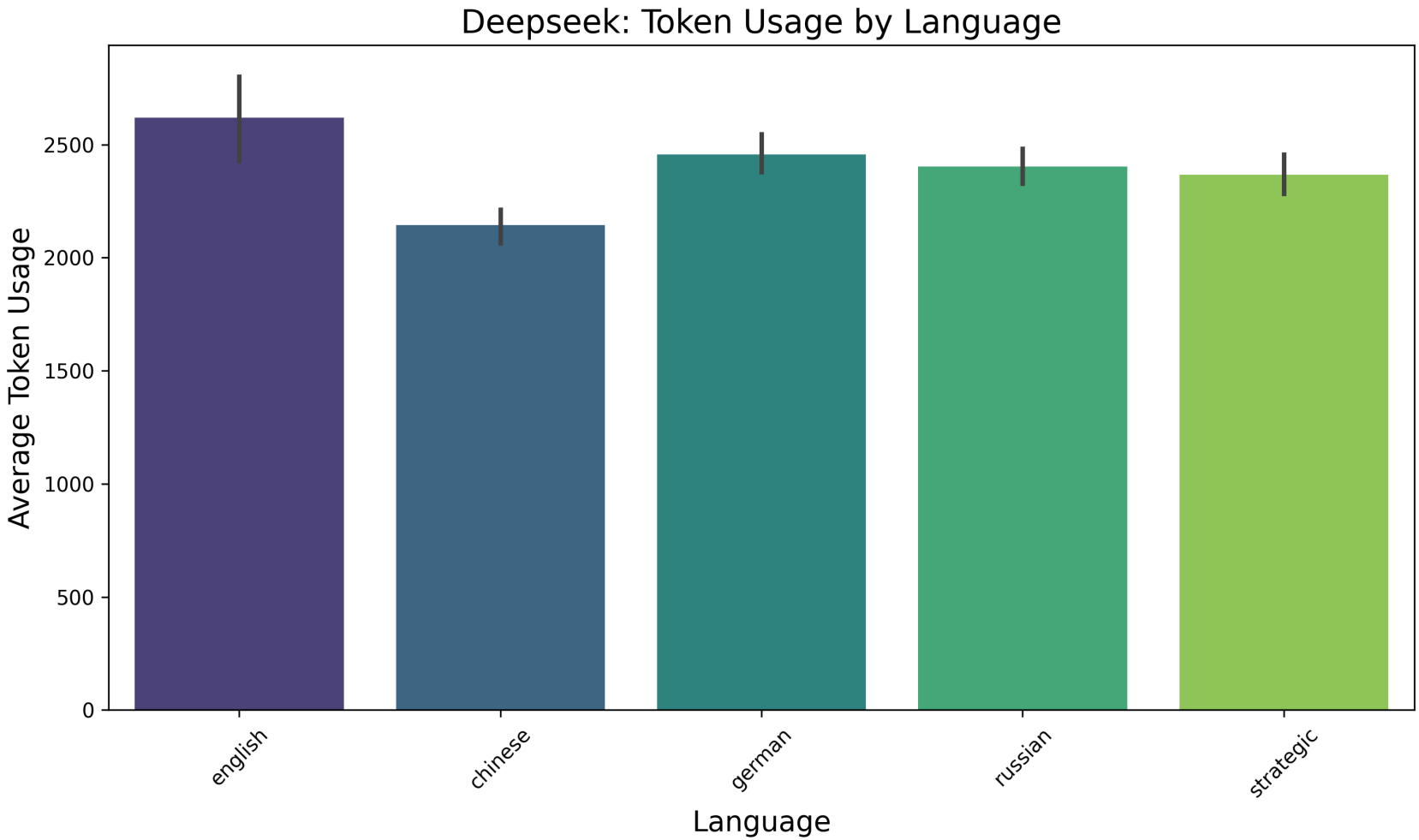
1. Classify problem type and difficulty
2. Select optimal language based on domain
3. Perform reasoning in selected language
4. Return answer in English

Claude: Token usage by language



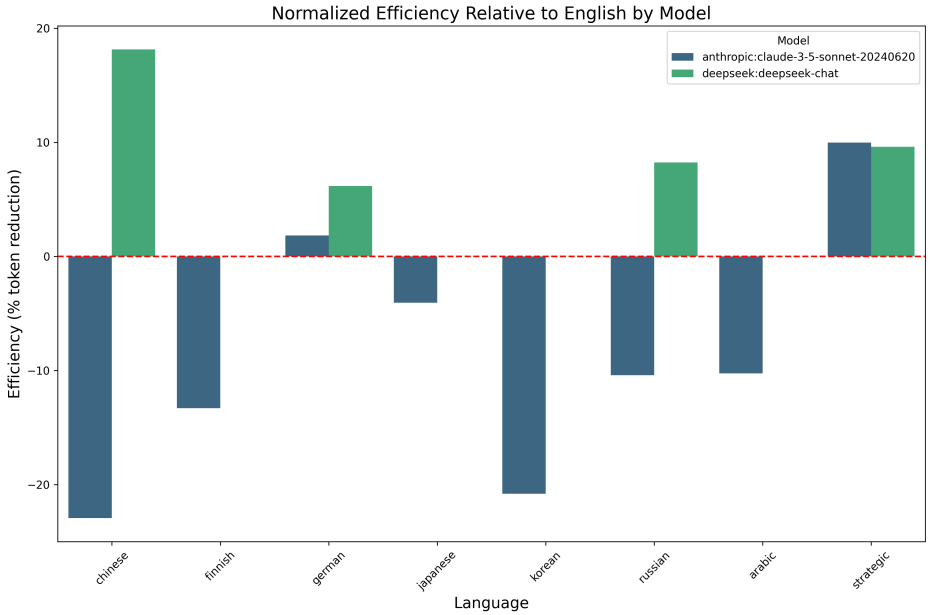
- Chinese and Korean have lowest token usages
- English has the highest token usage

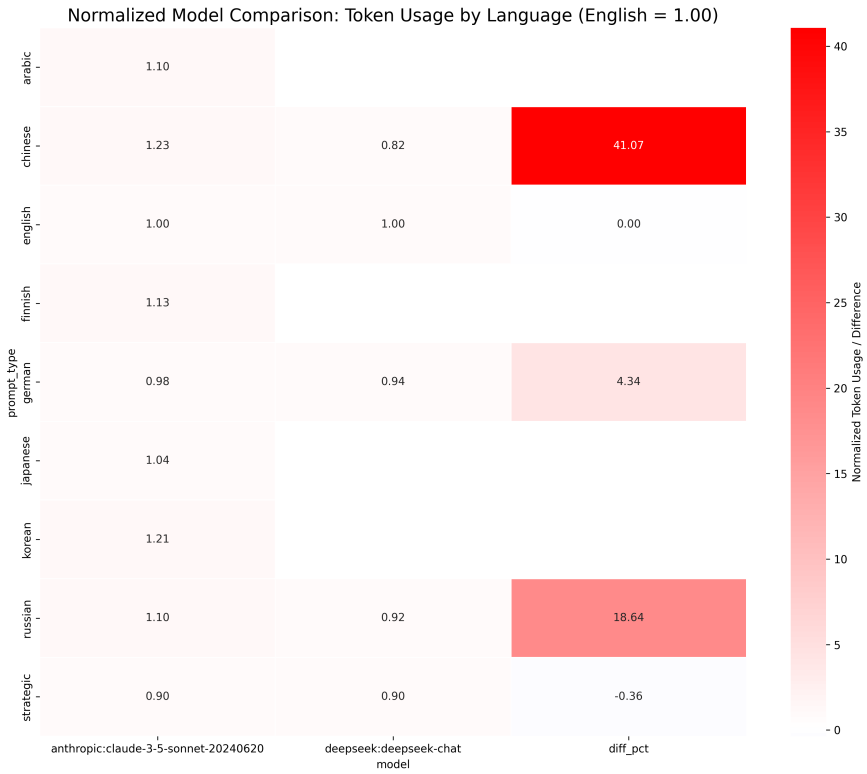
Deepseek: Token usage by language



- On average, Chinese maintains token efficiency
- English, still has the highest token usage

Model comparison: Claude vs. Deepseek





- Performance of Chinese differs depending on the model used.
- Strategic language selection maintains efficiency advantage.
- Deepseek shows higher efficiency gains for Chinese compared to Anthropic.

Model comparison: Claude vs. Deepseek

- Claude 3.7 Sonnet results available for all languages and benchmarks
- Deepseek Chat results available for select languages and long-context QA tasks
- Comparison shows differences in tokenization efficiency
- Chinese-developed models may have different efficiency patterns for Chinese text

Note

Results based on available data from successful API calls

Context Length Impact

- Long-context QA experiments have contexts ranging from 2,000 to 10,000+ characters
- English shows better efficiency for long contexts compared to other benchmarks, or languages
- Strategic language selection maintains efficiency advantage

Observation

The efficiency advantage of logographic systems diminishes as context length increases

Conclusion

Key observation

- Language efficiency for chain-of-thought reasoning varies significantly by domain
- Chinese excels at mathematical reasoning but underperforms in long-context tasks, showing diminishing advantages of logographic systems
- Strategic language selection could yield the highest overall efficiency

Next Steps

- Complete Deepseek model testing across all benchmarks
- Develop more sophisticated language selection algorithms
- Incorporate models primarily trained in different languages e.g. Qwen