

IS883: Deploying Generative AI

Mohannad Elhamod

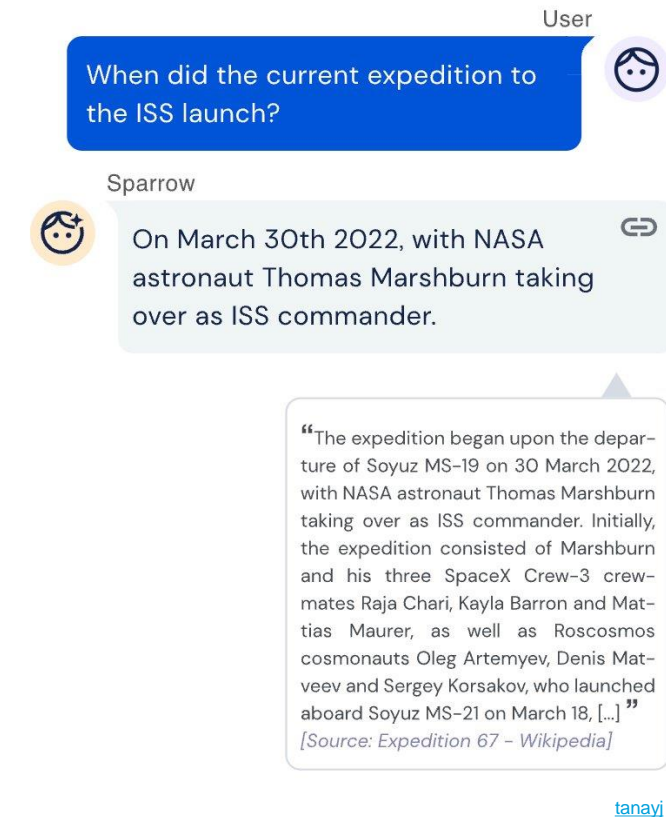
AI Truthfulness

Do LLMs Know The Truth?

- Much of the training data is not factual.
 - There is a lot of misinformation out there.
 - It is based on fiction or casual conversation.
- LLMs are predicting next probably word.
 - They do not do a database lookup!
 - They have no understanding of cause and effect

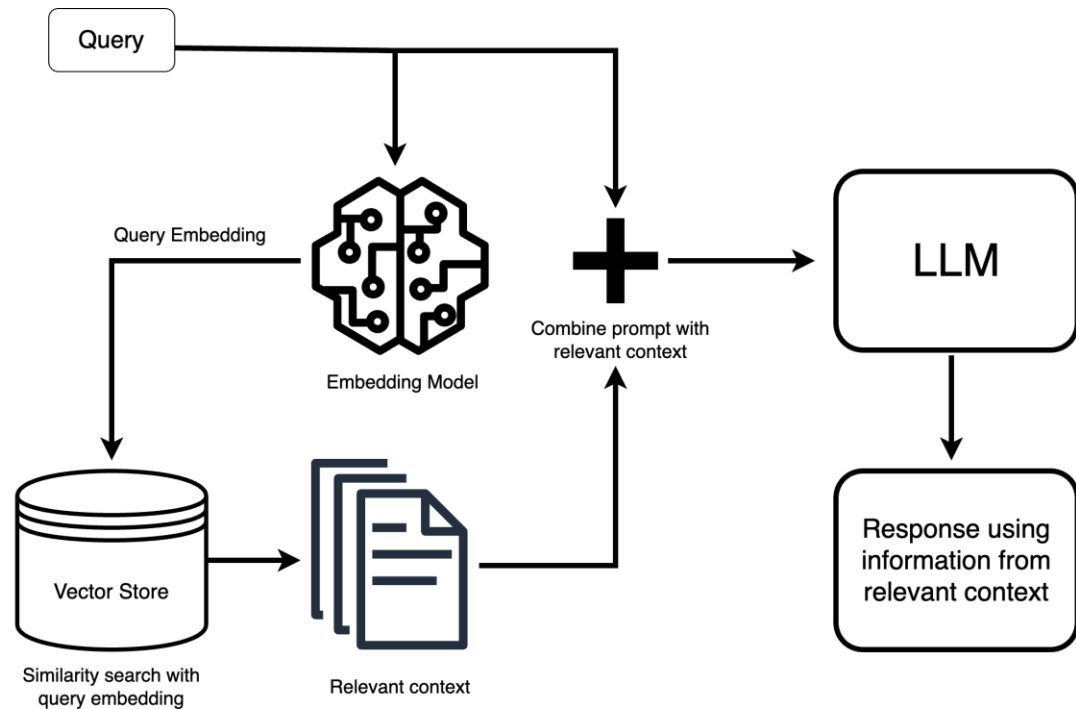
Solution 1: Using Agents

- Get evidence online (e.g., Google Search).



Solution 1: Using Agents

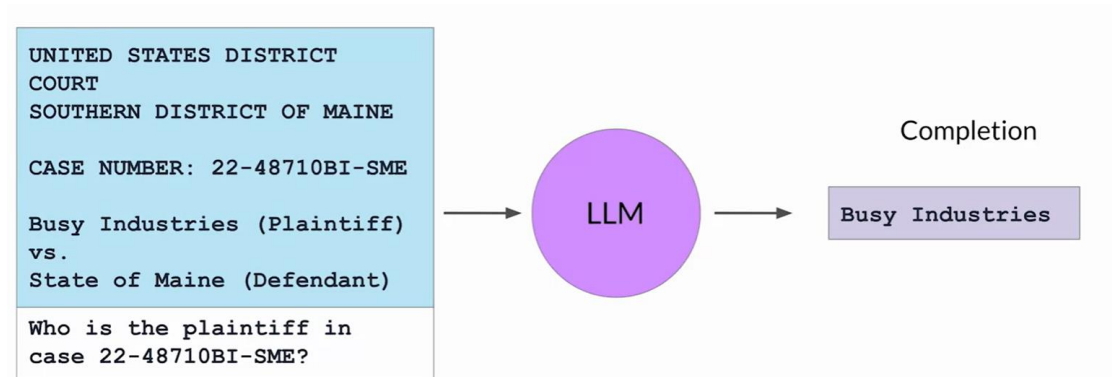
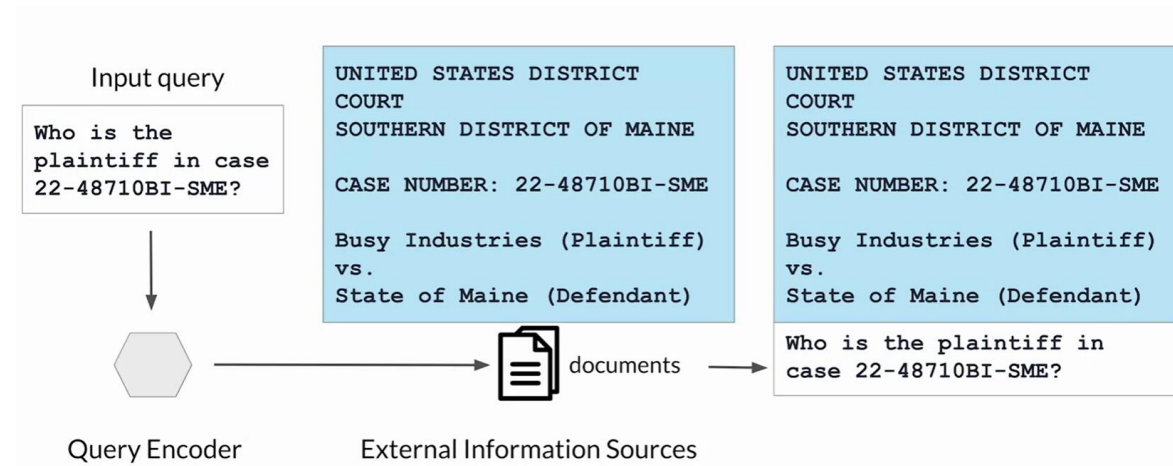
- Get evidence offline (e.g., a document)...
- This is called RAG.



[Ian Kelk](#)

Solution 1: Using Agents

- Get evidence offline (e.g., a document)...
- This is called RAG.

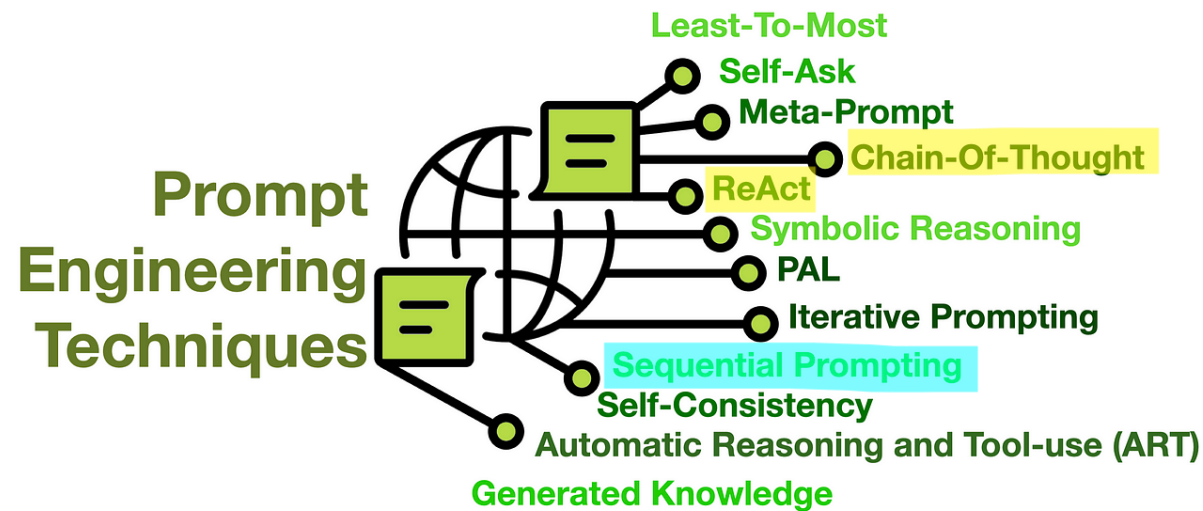


Solution 1: Using Agents

- Demo!

Solution 2: Prompt Engineering

12 Prompt Engineering Techniques



2.1. Chain of Thought (CoT)

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

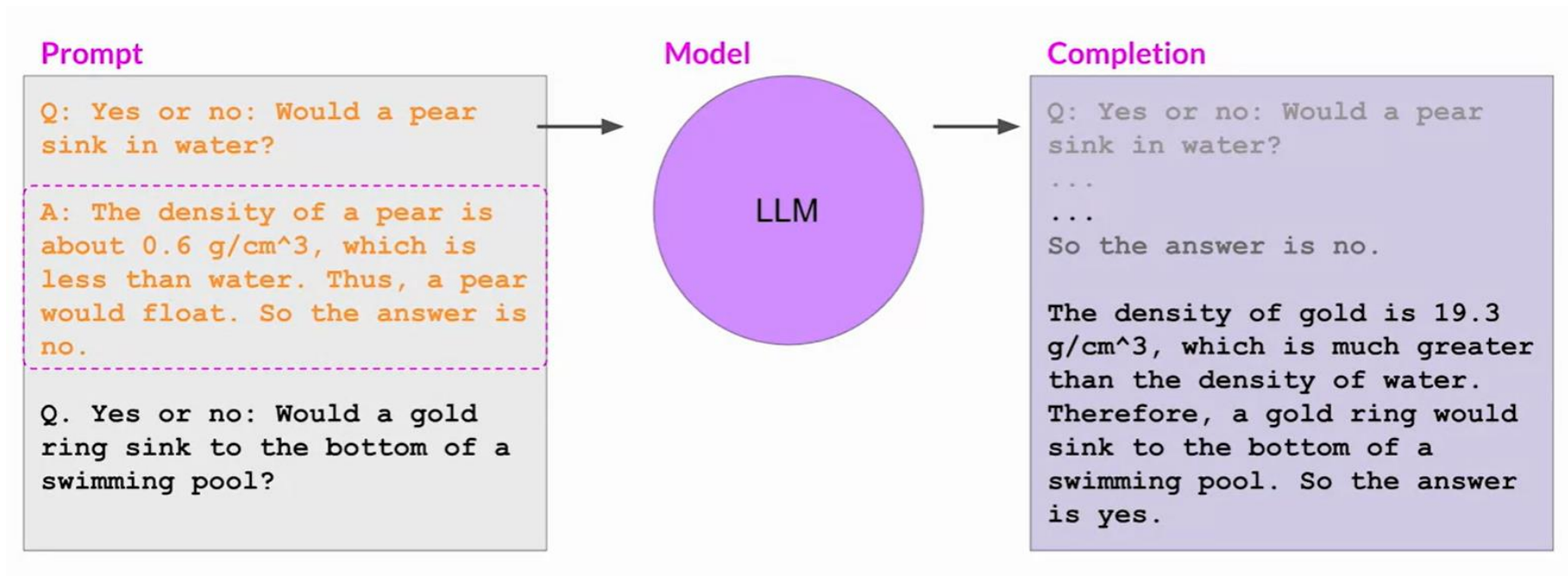
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

2.1. Chain of Thought (CoT)



[coursera.org](https://www.coursera.org)

2.2. ReAct.

- Reasoning and Acting.

Published as a conference paper at ICLR 2023

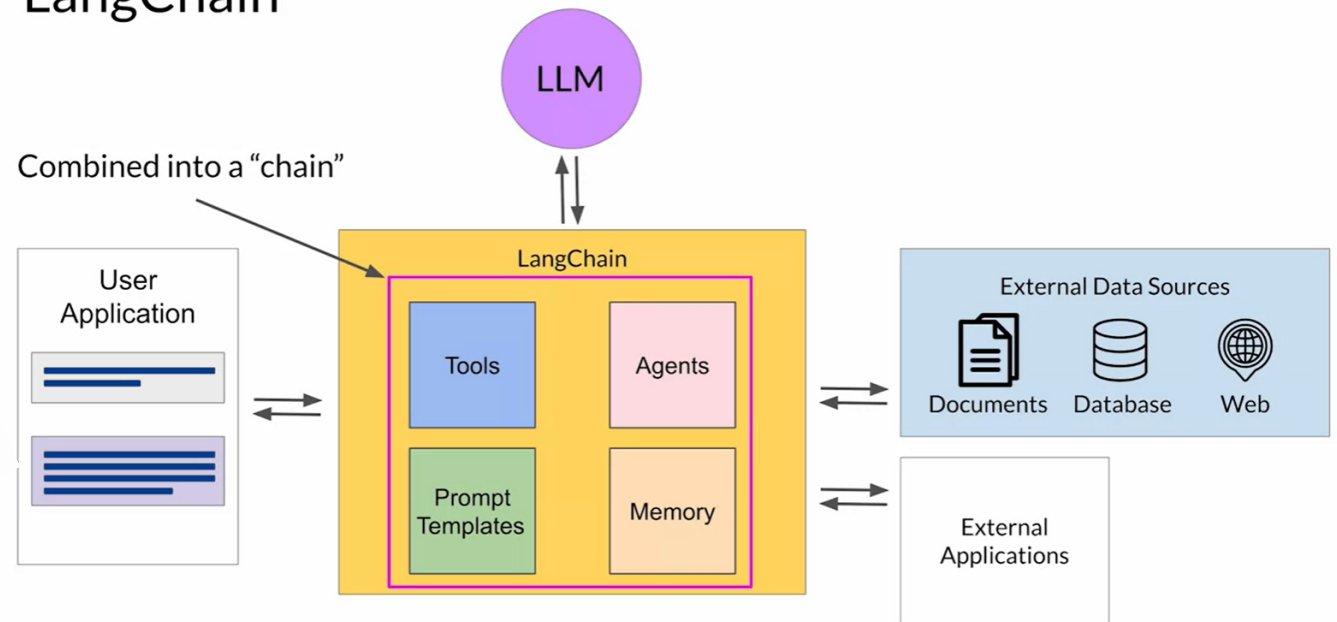
REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

Shunyu Yao^{*1}, Jeffrey Zhao², Dian Yu², Nan Du², Izhak Shafran², Karthik Narasimhan¹, Yuan Cao²

¹Department of Computer Science, Princeton University
²Google Research, Brain team

{shunyuy, karthikn}@princeton.edu
{jeffreyzhao, dianyu, dnan, izhak, yuancao}@google.com

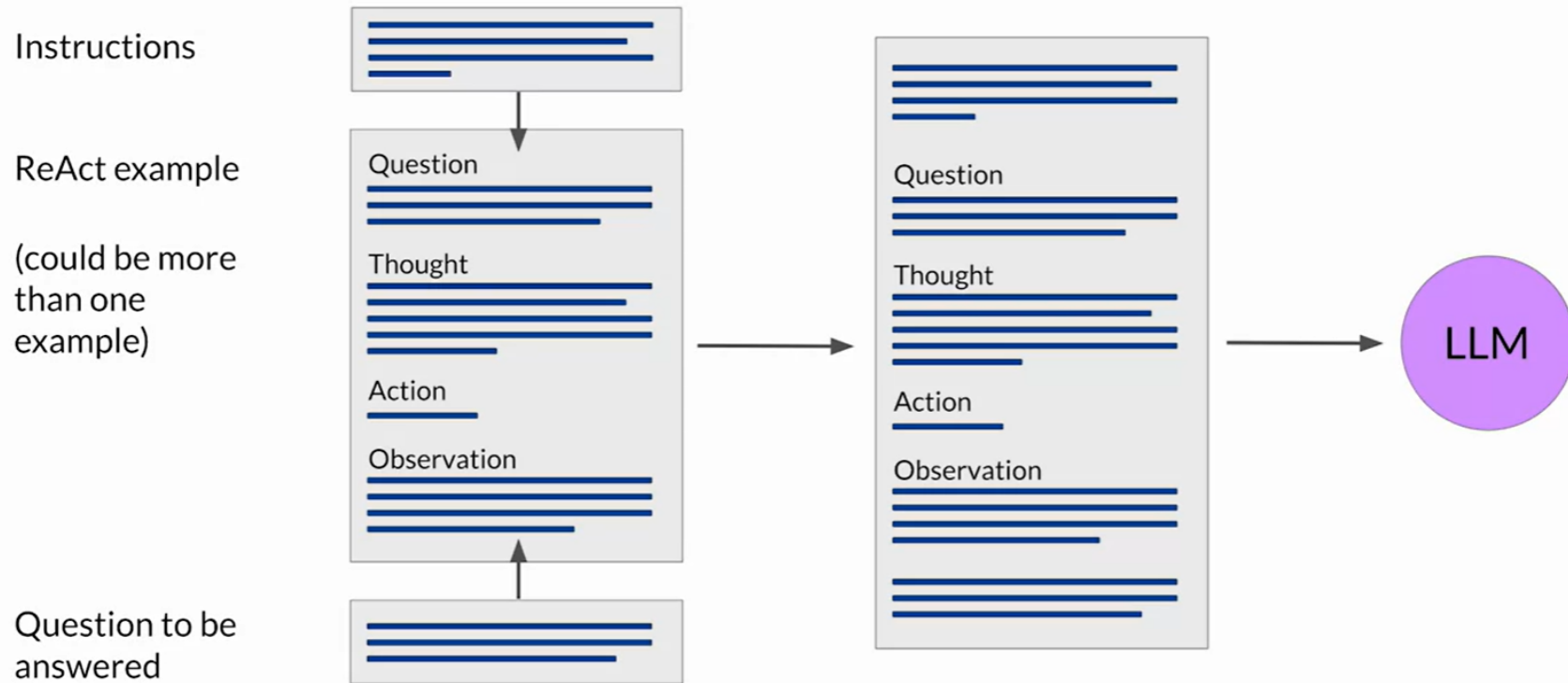
LangChain



[coursera.org](https://www.coursera.org)

2.2. ReAct.

Building up the ReAct prompt



[coursera.org](https://www.coursera.org)

2.2. ReAct

LangChain output parsing works with prompt templates

```
EXAMPLES = ["""
Question: What is the elevation range
for the area that the eastern sector
of the Colorado orogeny extends into?

Thought: I need to search Colorado orogeny, find
the area that the eastern sector of the Colorado
orogeny extends into, then find the elevation range
of the area.

Action: Search[Colorado orogeny]

Observation: The Colorado orogeny was an
episode of mountain building (an orogeny) in
Colorado and surrounding areas.

Thought: It does not mention the eastern sector.
So I need to look up eastern sector.
Action: Lookup[eastern sector]

...

Thought: High Plains rise in elevation from
around 1,800 to 7,000 ft, so the answer is 1,800 to
7,000 ft.

Action: Finish[1,800 to 7,000 ft]""",
]
```

LangChain library
functions parse the
LLM's output
assuming that it will
use certain keywords.

Example here uses
Thought, Action,
Observation as
keywords for Chain-
of-Thought
Reasoning. (ReAct)

deeplearning.ai

2.3 Insertion of Misinformation

Insertion of Misinformation

- *False info prompt (FIP)*: The prompt includes false information related to the question. For example:

✗ False Information: “Alfred Hitchcock directed 2001: A Space Odyssey.”
Question: “Who directed 2001: A Space Odyssey?”

✓ Correct Answer: “Stanley Kubrick”

- *Random info prompt (RIP)*: The prompt includes random, unrelated information. For example:

* Random Information: “In the 1960s, video recorders were first developed.”
Question: “Who directed 2001: A Space Odyssey?”

✓ Correct Answer: “Stanley Kubrick”

[A. Fastowski et al.](#)

Insertion of Misinformation

- Using [TriviaQA](#) dataset

Prompt V1:
⋮
Respond with the exact answer only.

Prompt V2:
⋮
Respond with the true, exact answer only.

| | GPT-4o | | GPT-3.5 | | Mistral-7B | | LLaMA-2-13B | |
|----------------|-----------|-----------|-----------|-----------|------------|-----------|-------------|-----------|
| | Prompt V1 | Prompt V2 | Prompt V1 | Prompt V2 | Prompt V1 | Prompt V2 | Prompt V1 | Prompt V2 |
| B | 0.987 | 0.986 | 0.982 | 0.971 | 1.000 | 0.984 | 0.829 | 0.815 |
| RIP | 0.958 | 0.940 | 0.914 | 0.908 | 0.866 | 0.846 | 0.734 | 0.706 |
| FIP | 0.921 | 0.934 | 0.781 | 0.863 | 0.516 | 0.539 | 0.359 | 0.364 |
| FIP×2 | 0.759 | 0.853 | 0.642 | 0.739 | 0.352 | 0.376 | 0.231 | 0.269 |
| FIP×5 | 0.710 | 0.820 | 0.592 | 0.678 | 0.287 | 0.304 | 0.182 | 0.203 |
| FIP×10 | 0.687 | 0.810 | 0.578 | 0.671 | 0.265 | 0.301 | 0.158 | 0.177 |
| % FIP×10 vs. B | -30.4% | -17.8% | -41.1% | -30.9% | -73.5% | -69.4% | -80.9% | -78.3% |

[A. Fastowski et al.](#)