FLAME🔥: Factuality-Aware Alignment for Large Language Models

Presenter: Sheng-Chieh (Jack) Lin, s269lin@uwaterloo.ca

Sheng-Chieh Lin*[1], Luyu Gao[2], Barlas Oguz[3], Wenhan Xiong[3], Jimmy Lin[1], Wen-tau Yih[3], and Xilun Chen*[3]
University of Waterloo[1], Carnegie Mellon University[2], Meta AI[3]
s269lin@uwaterloo.ca, xilun@meta.com

What Yields LLMs' Hallucinations?

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.

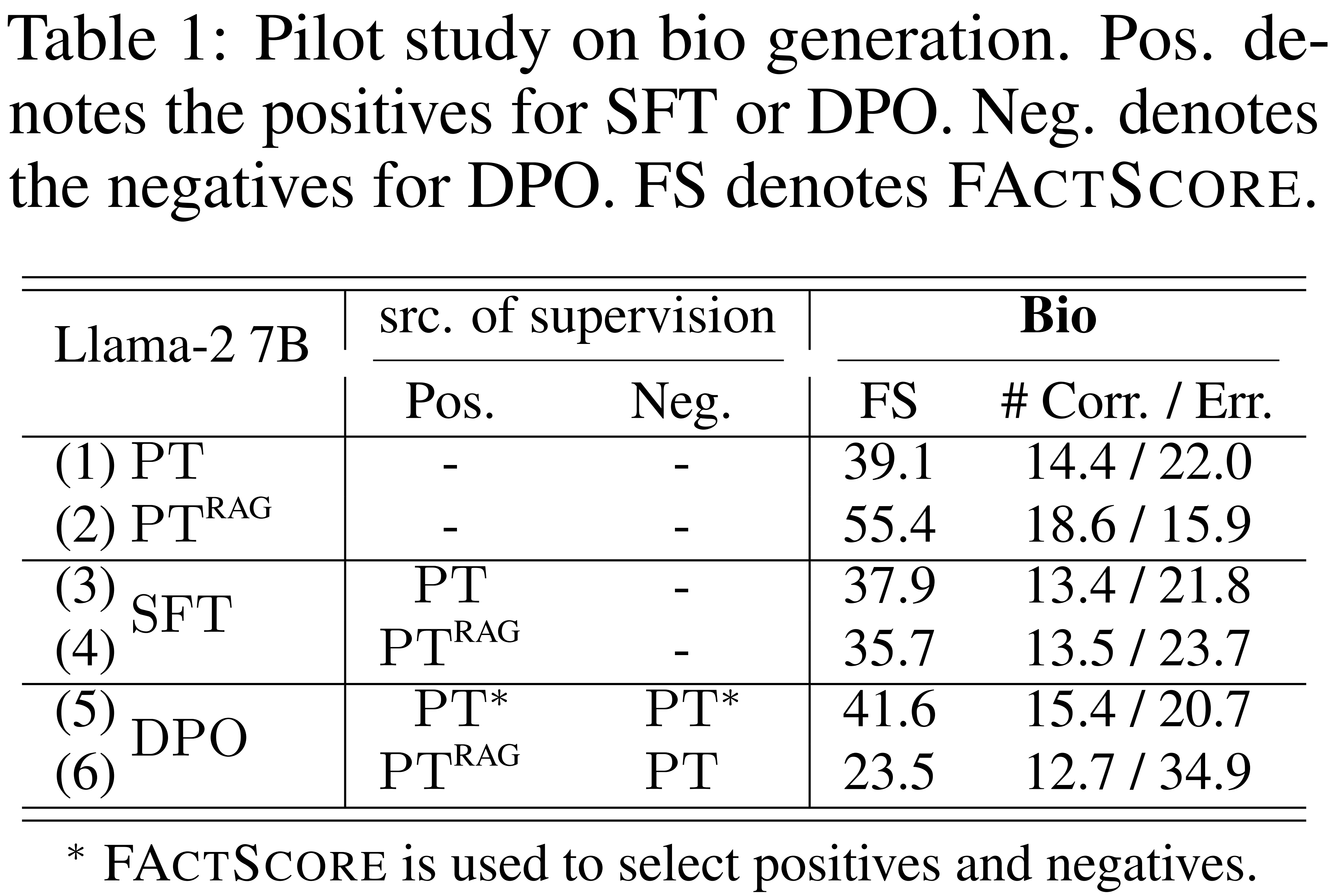


Table 1: Pilot study on bio generation. Pos. denotes the positives for SFT or DPO. Neg. denotes the negatives for DPO. FS denotes FActScore.

Llama-2 7B	src. of supervision		Bio	
	Pos.	Neg.	FS	# Corr. / Err.
(1) PT	-	-	39.1	14.4 / 22.0
(2) PT^{RAG}	-	-	55.4	18.6 / 15.9
(3) SFT	PT	-	37.9	13.4 / 21.8
(4)	PT^{RAG}	-	35.7	13.5 / 23.7
(5) DPO	PT*	PT*	41.6	15.4 / 20.7
(6)	PT^{RAG}	PT	23.5	12.7 / 34.9

* FActScore is used to select positives and negatives.

What Yields LLMs' Hallucinations?

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.

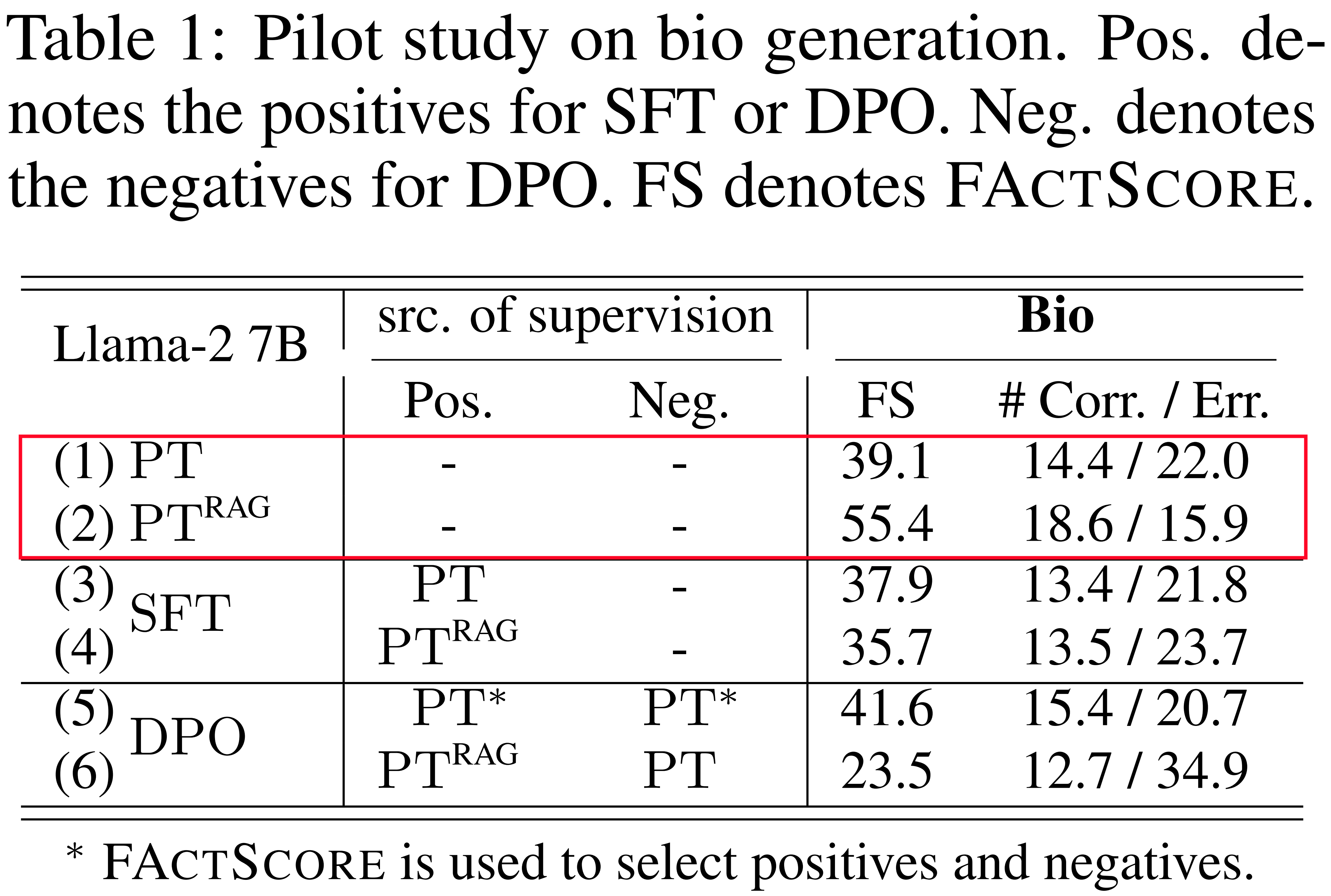


Table 1: Pilot study on bio generation. Pos. denotes the positives for SFT or DPO. Neg. denotes the negatives for DPO. FS denotes FactScore.

Llama-2 7B	src. of supervision		Bio	
	Pos.	Neg.	FS	# Corr. / Err.
(1) PT	-	-	39.1	14.4 / 22.0
(2) PT^{RAG}	-	-	55.4	18.6 / 15.9
(3) SFT	PT	-	37.9	13.4 / 21.8
(4) SFT	PT^{RAG}	-	35.7	13.5 / 23.7
(5) DPO	PT*	PT*	41.6	15.4 / 20.7
(6) DPO	PT^{RAG}	PT	23.5	12.7 / 34.9

* FactScore is used to select positives and negatives.

What Yields LLMs' Hallucinations?

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.

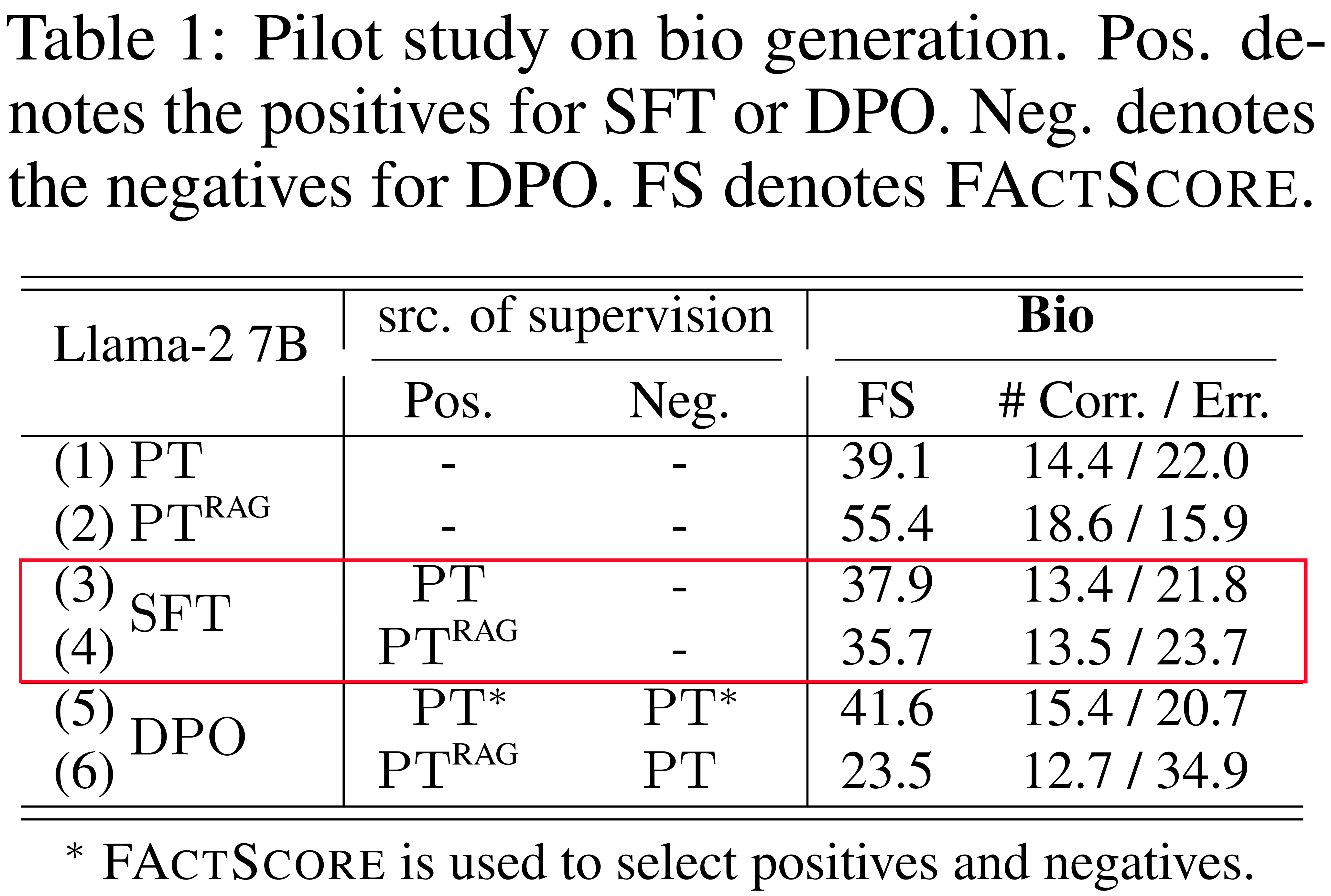


Table 1: Pilot study on bio generation. Pos. denotes the positives for SFT or DPO. Neg. denotes the negatives for DPO. FS denotes FACTSCORE.

Llama-2 7B	src. of supervision		Bio	
	Pos.	Neg.	FS	# Corr. / Err.
(1) PT	-	-	39.1	14.4 / 22.0
(2) PT^{RAG}	-	-	55.4	18.6 / 15.9
(3) SFT	PT	-	37.9	13.4 / 21.8
(4)	PT^{RAG}	-	35.7	13.5 / 23.7
(5) DPO	PT*	PT*	41.6	15.4 / 20.7
(6)	PT^{RAG}	PT	23.5	12.7 / 34.9

* FACTSCORE is used to select positives and negatives.

What Yields LLMs' Hallucinations?

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.

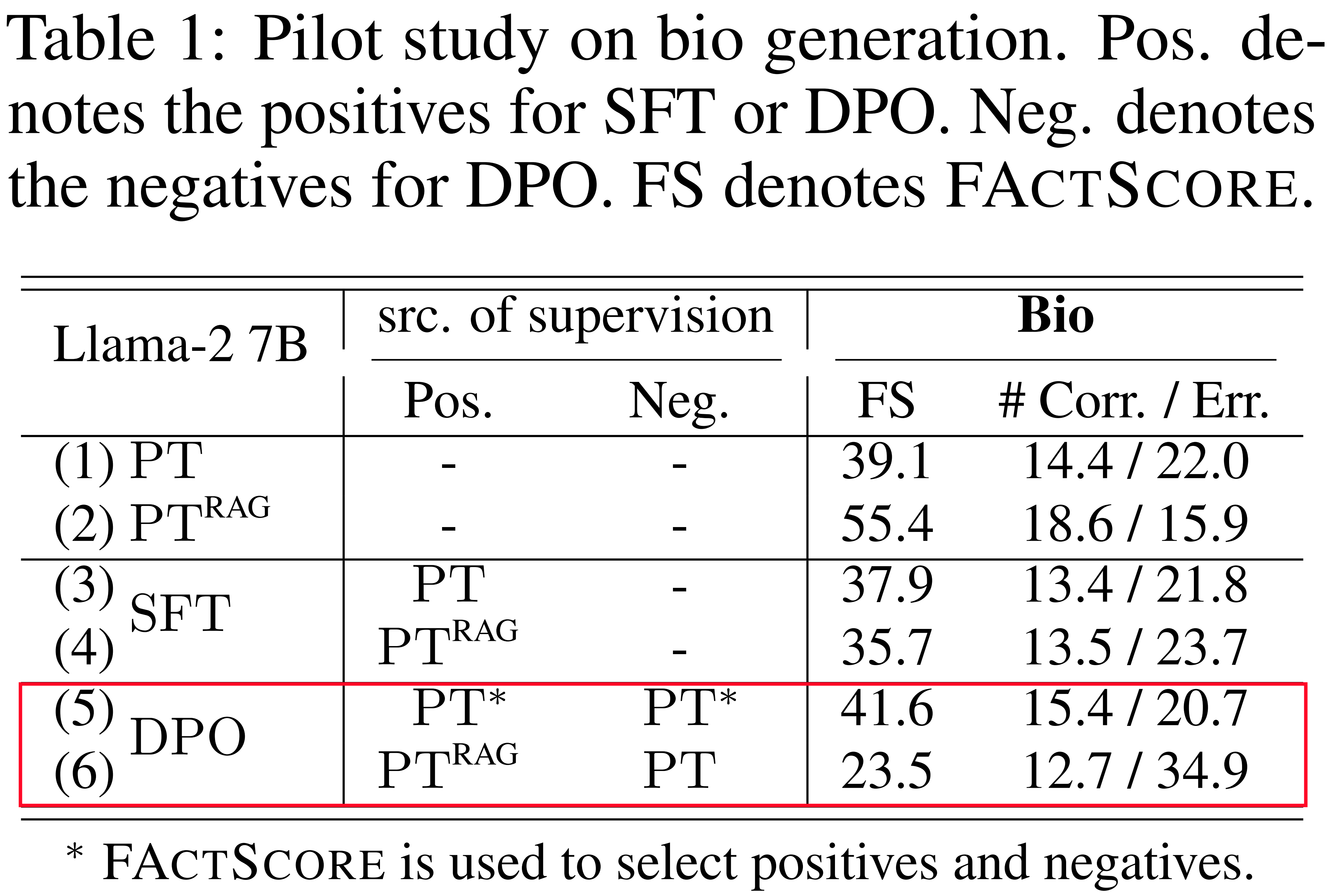


Table 1: Pilot study on bio generation. Pos. denotes the positives for SFT or DPO. Neg. denotes the negatives for DPO. FS denotes FACTSCORE.

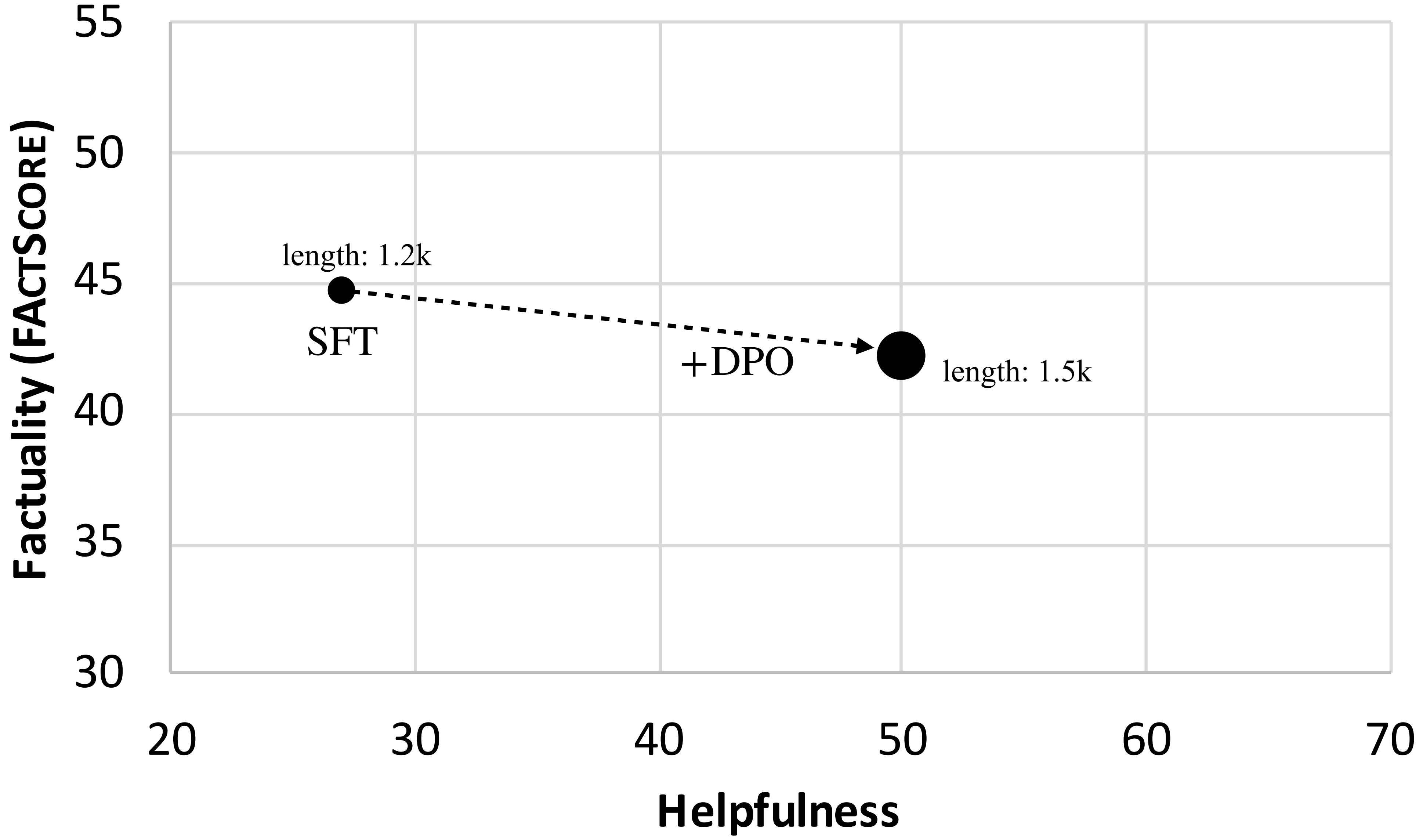
Llama-2 7B	src. of supervision		**Bio**	
	Pos.	Neg.	FS	# Corr. / Err.
(1) PT	-	-	39.1	14.4 / 22.0
(2) PT^{RAG}	-	-	55.4	18.6 / 15.9
(3) SFT	PT	-	37.9	13.4 / 21.8
(4) SFT	PT^{RAG}	-	35.7	13.5 / 23.7
(5) DPO	PT*	PT*	41.6	15.4 / 20.7
(6) DPO	PT^{RAG}	PT	23.5	12.7 / 34.9

* FACTSCORE is used to select positives and negatives.

What Yields LLMs' Hallucinations?

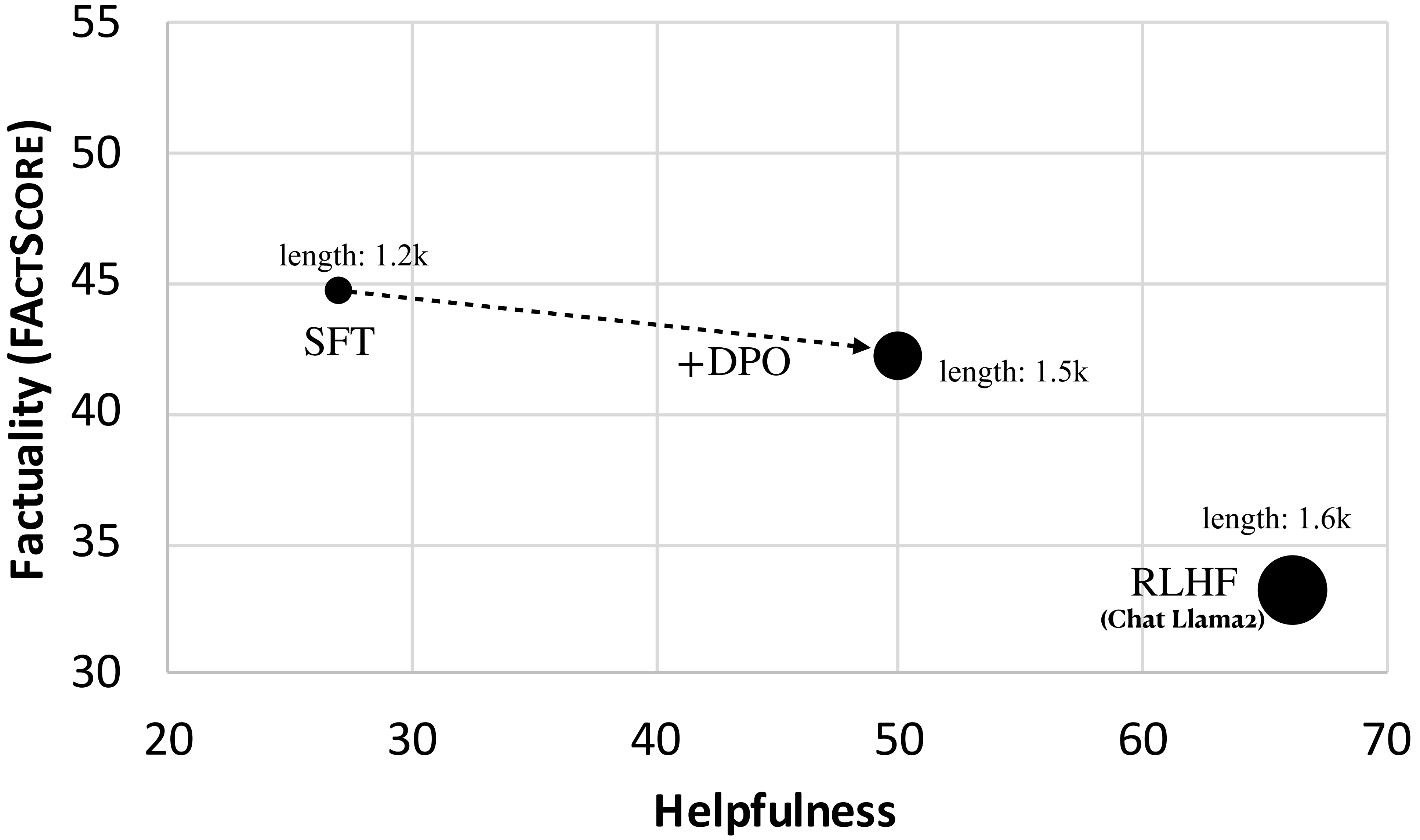
- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.

- Observation2: Instruction following alignment often guides LLMs to output more detailed and longer responses, but inevitably yields more hallucinations.



What Yields LLMs' Hallucinations?

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.
- Observation2: Instruction following alignment often guides LLMs to output more detailed and longer responses, but inevitably yields more hallucinations.



Hypothetical Root Causes

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.
- Cause: Supervised fine-tuning (SFT) may force LLMs to learn unknown knowledge.

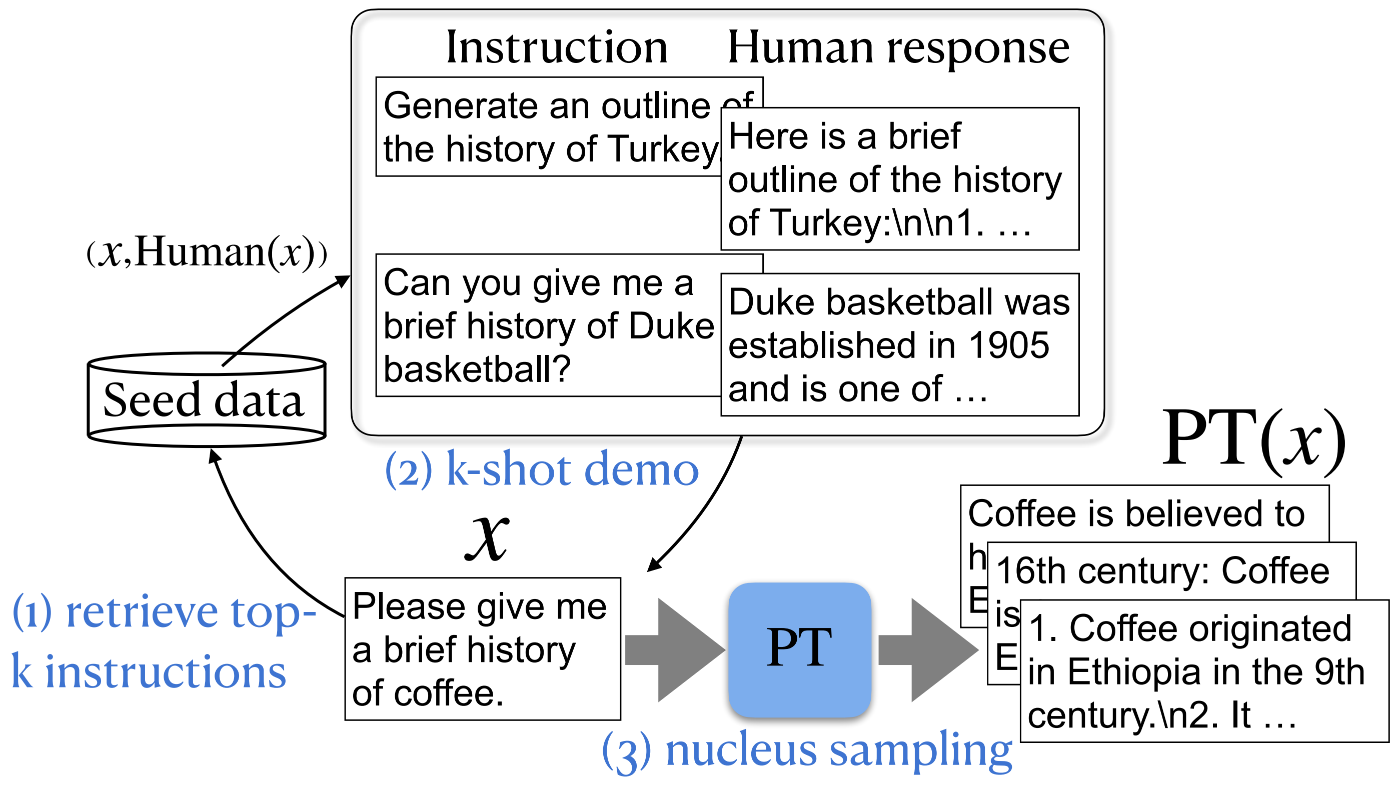
Hypothetical Root Causes

- Observation1: Supervised fine-tuning (SFT) LLMs with more factual responses does not necessarily yield LLMs with better factuality.
- Cause: Supervised fine-tuning (SFT) may force LLMs to learn unknown knowledge.
- Observation2: Instruction following alignment often guides LLMs to output more detailed and longer responses, but inevitably yields more hallucinations.
- Cause: Reward models at RL stage encourage to output detailed responses but ignore factuality[1].

[1] Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In Proc. ICLR, 2024.

Proposed Solutions

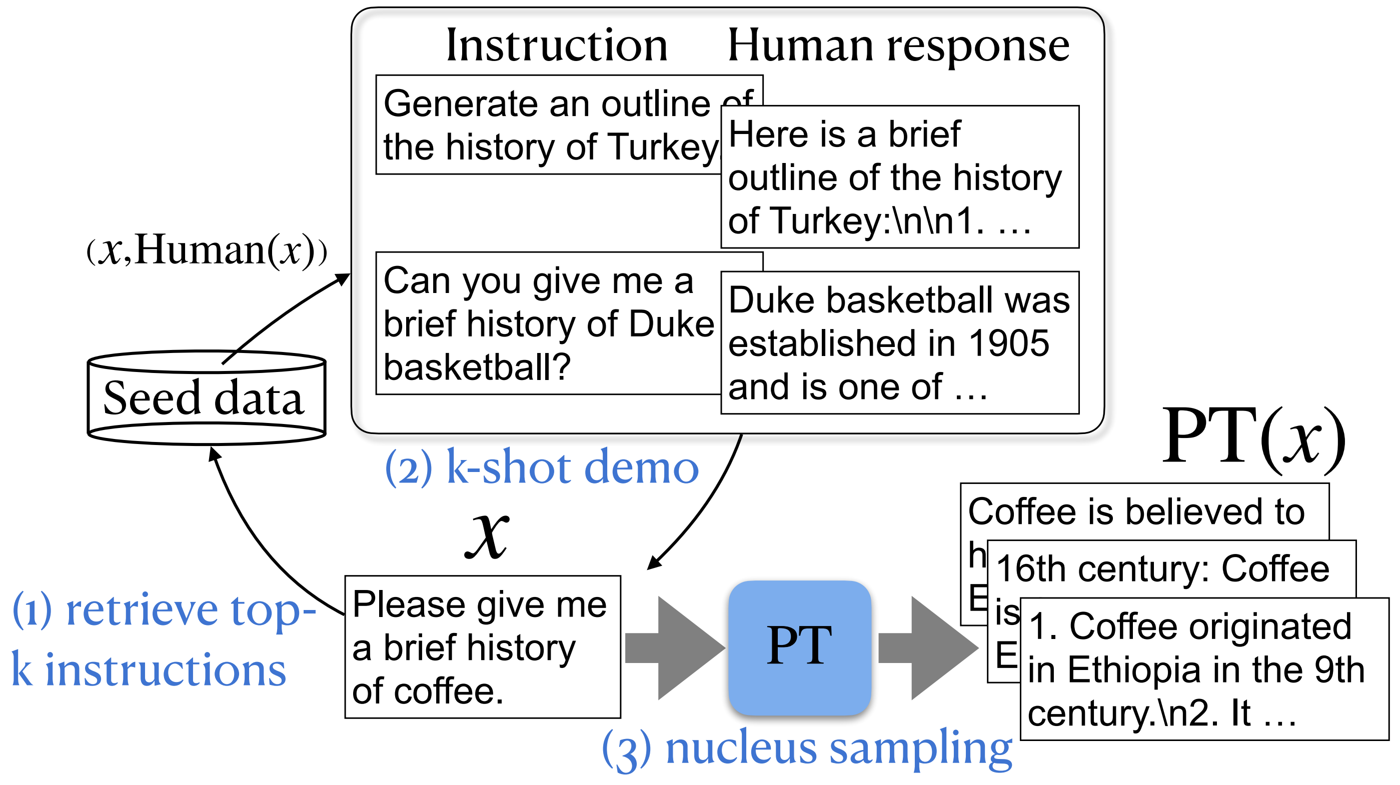
- Factuality SFT: instead of using human written responses, we prompt pre-trained LLMs to output responses as self supervisions to avoid unfamiliar knowledge.


Instruction
Human response
Generate an outline of the history of Turkey
Here is a brief outline of the history of Turkey:\n\n1. …
Can you give me a brief history of Duke basketball?
Duke basketball was established in 1905 and is one of …
$(x, \text{Human}(x))$
Seed data
(2) k-shot demo
x
(1) retrieve top-k instructions
Please give me a brief history of coffee.
PT
(3) nucleus sampling
$\text{PT}(x)$
Coffee is believed to
16th century: Coffee
1. Coffee originated in Ethiopia in the 9th century.\n2. It …

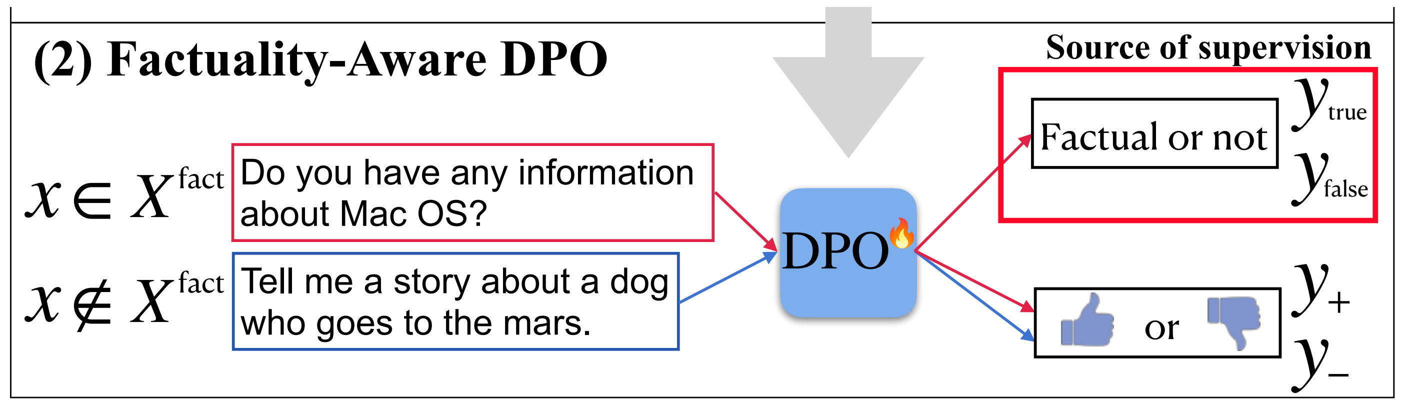
(a)

Proposed Solutions

- Factuality SFT: instead of using human written responses, we prompt pre-trained LLMs to output responses as self supervisions to avoid unfamiliar knowledge.
- Factuality DPO: we explicitly create preference pairs with factuality reward models.


Instruction
Human response
Generate an outline of the history of Turkey
Here is a brief outline of the history of Turkey:\n\n1. …
Can you give me a brief history of Duke basketball?
Duke basketball was established in 1905 and is one of …
(x,Human(x))
Seed data
(2) k-shot demo
x
(1) retrieve top-k instructions
Please give me a brief history of coffee.
PT
(3) nucleus sampling
PT(x)
Coffee is believed to
16th century: Coffee
1. Coffee originated in Ethiopia in the 9th century.\n2. It …

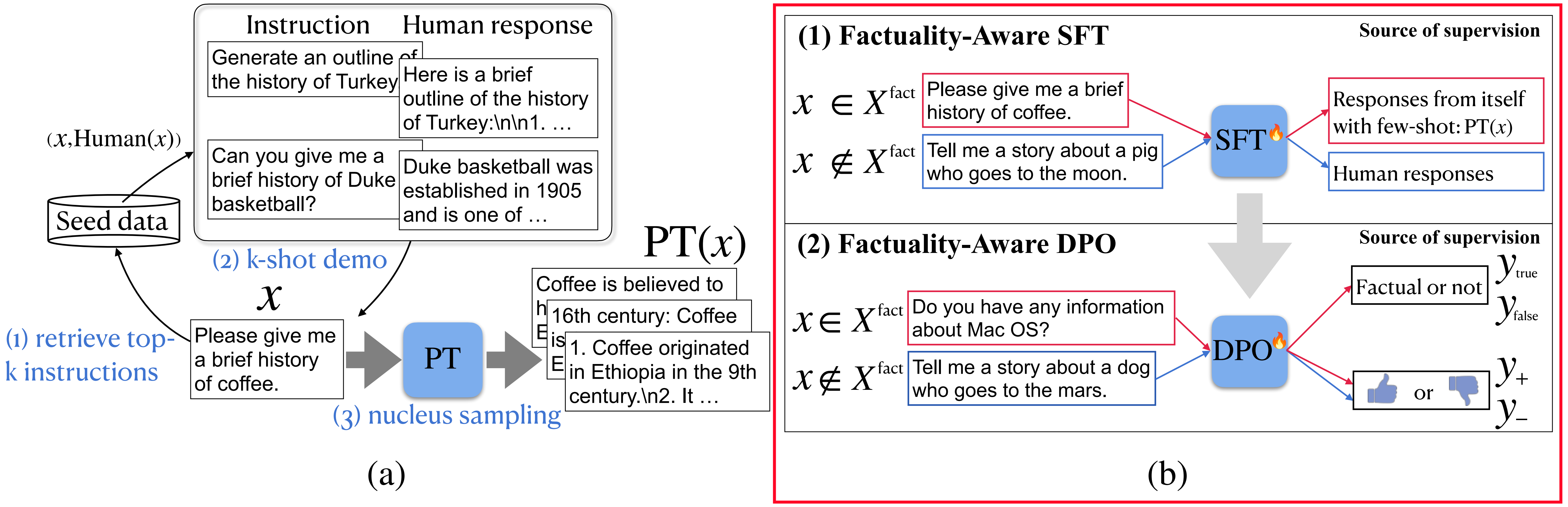
(a)



(b)

Proposed Solutions

- Factuality SFT: instead of using human written responses, we prompt pre-trained LLMs to output responses as self supervisions to avoid unfamiliar knowledge.
- Factuality DPO: we explicitly create preference pairs with factuality reward models.
- FLAME🔥: In addition to conducting SFT and DPO for instruction following, fact-based instructions are detected for factuality SFT and DPO.

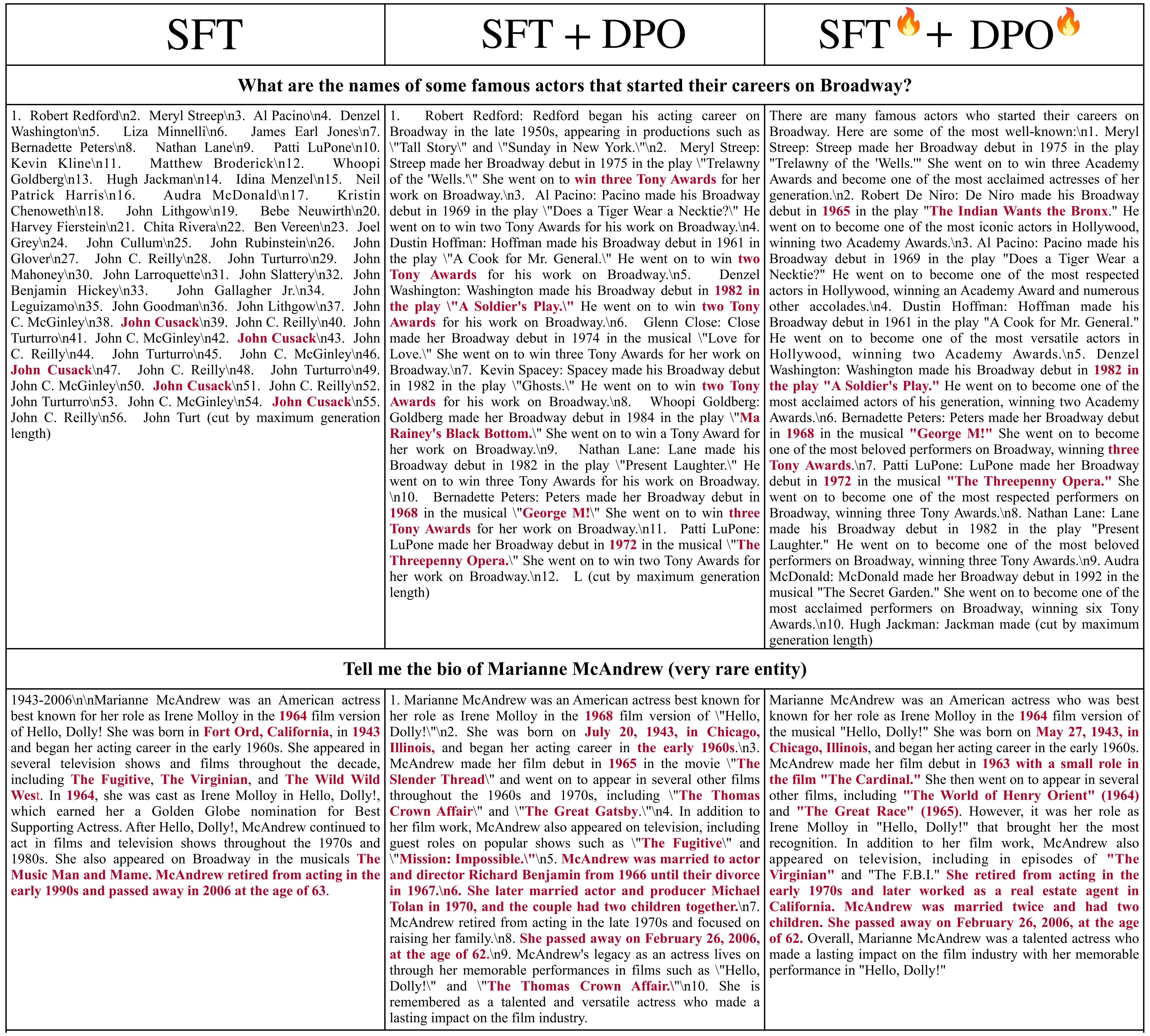

Instruction
Human response
Generate an outline of the history of Turkey
Here is a brief outline of the history of Turkey:\n\n1. …
Can you give me a brief history of Duke basketball?
Duke basketball was established in 1905 and is one of …
(x,Human(x))
Seed data
(2) k-shot demo
x
(1) retrieve top-k instructions
Please give me a brief history of coffee.
PT
(3) nucleus sampling
PT(x)
Coffee is believed to
16th century: Coffee
1. Coffee originated in Ethiopia in the 9th century.\n2. It …
(1) Factuality-Aware SFT
Source of supervision
$x \in X^{\text{fact}}$
Please give me a brief history of coffee.
$x \notin X^{\text{fact}}$
Tell me a story about a pig who goes to the moon.
SFT
Responses from itself with few-shot: PT(x)
Human responses
(2) Factuality-Aware DPO
Source of supervision
$x \in X^{\text{fact}}$
Do you have any information about Mac OS?
$x \notin X^{\text{fact}}$
Tell me a story about a dog who goes to the mars.
DPO
Factual or not
y_{true}
y_{false}
or
y_+
y_-

(a)

(b)

Qualitative Analysis

- SFT + DPO tends to generate more lengthy and detailed responses regardless of the questions.
- FLAEME automatically adjusts the length of the responses according to the difficulty of questions.



SFT	SFT + DPO	SFT🔥 + DPO🔥
What are the names of some famous actors that started their careers on Broadway?		
1. Robert Redford\n2. Meryl Streep\n3. Al Pacino\n4. Denzel Washington\n5. Liza Minnelli\n6. James Earl Jones\n7. Bernadette Peters\n8. Nathan Lane\n9. Patti LuPone\n10. Kevin Kline\n11. Matthew Broderick\n12. Whoopi Goldberg\n13. Hugh Jackman\n14. Idina Menzel\n15. Neil Patrick Harris\n16. Audra McDonald\n17. Kristin Chenoweth\n18. John Lithgow\n19. Bebe Neuwirth\n20. Harvey Fierstein\n21. Chita Rivera\n22. Ben Vereen\n23. Joel Grey\n24. John Cullum\n25. John Rubinstein\n26. John Glover\n27. John C. Reilly\n28. John Turturro\n29. John Mahoney\n30. John Larroquette\n31. John Slattery\n32. John Benjamin Hickey\n33. John Gallagher Jr.\n34. John Leguizamo\n35. John Goodman\n36. John Lithgow\n37. John C. McGinley\n38. **John Cusack**\n39. John C. Reilly\n40. John Turturro\n41. John C. McGinley\n42. **John Cusack**\n43. John C. Reilly\n44. John Turturro\n45. John C. McGinley\n46. **John Cusack**\n47. John C. Reilly\n48. John Turturro\n49. John C. McGinley\n50. **John Cusack**\n51. John C. Reilly\n52. John Turturro\n53. John C. McGinley\n54. **John Cusack**\n55. John C. Reilly\n56. John Turt (cut by maximum generation length)	1. Robert Redford: Redford began his acting career on Broadway in the late 1950s, appearing in productions such as \"Tall Story\" and \"Sunday in New York.\"\n2. Meryl Streep: Streep made her Broadway debut in 1975 in the play \"Trelawny of the 'Wells.'\" She went on to **win three Tony Awards** for her work on Broadway.\n3. Al Pacino: Pacino made his Broadway debut in 1969 in the play \"Does a Tiger Wear a Necktie?\" He went on to win two Tony Awards for his work on Broadway.\n4. Dustin Hoffman: Hoffman made his Broadway debut in 1961 in the play \"A Cook for Mr. General.\" He went on to win **two Tony Awards** for his work on Broadway.\n5. Denzel Washington: Washington made his Broadway debut in **1982 in the play \"A Soldier's Play.\"** He went on to win **two Tony Awards** for his work on Broadway.\n6. Glenn Close: Close made her Broadway debut in 1974 in the musical \"Love for Love.\" She went on to win three Tony Awards for her work on Broadway.\n7. Kevin Spacey: Spacey made his Broadway debut in 1982 in the play \"Ghosts.\" He went on to win **two Tony Awards** for his work on Broadway.\n8. Whoopi Goldberg: Goldberg made her Broadway debut in 1984 in the play \"**Ma Rainey's Black Bottom.**\" She went on to win a Tony Award for her work on Broadway.\n9. Nathan Lane: Lane made his Broadway debut in 1982 in the play \"Present Laughter.\" He went on to win three Tony Awards for his work on Broadway. \n10. Bernadette Peters: Peters made her Broadway debut in **1968** in the musical \"**George M!**\" She went on to win **three Tony Awards** for her work on Broadway.\n11. Patti LuPone: LuPone made her Broadway debut in **1972** in the musical \"**The Threepenny Opera.**\" She went on to win two Tony Awards for her work on Broadway.\n12. L (cut by maximum generation length)	There are many famous actors who started their careers on Broadway. Here are some of the most well-known:\n1. Meryl Streep: Streep made her Broadway debut in 1975 in the play "Trelawny of the 'Wells.'" She went on to win three Academy Awards and become one of the most acclaimed actresses of her generation.\n2. Robert De Niro: De Niro made his Broadway debut in **1965** in the play "**The Indian Wants the Bronx**." He went on to become one of the most iconic actors in Hollywood, winning two Academy Awards.\n3. Al Pacino: Pacino made his Broadway debut in 1969 in the play "Does a Tiger Wear a Necktie?" He went on to become one of the most respected actors in Hollywood, winning an Academy Award and numerous other accolades.\n4. Dustin Hoffman: Hoffman made his Broadway debut in 1961 in the play "A Cook for Mr. General." He went on to become one of the most versatile actors in Hollywood, winning two Academy Awards.\n5. Denzel Washington: Washington made his Broadway debut in **1982 in the play "A Soldier's Play."** He went on to become one of the most acclaimed actors of his generation, winning two Academy Awards.\n6. Bernadette Peters: Peters made her Broadway debut in **1968** in the musical **"George M!"** She went on to become one of the most beloved performers on Broadway, winning **three Tony Awards**.\n7. Patti LuPone: LuPone made her Broadway debut in **1972** in the musical **"The Threepenny Opera."** She went on to become one of the most respected performers on Broadway, winning three Tony Awards.\n8. Nathan Lane: Lane made his Broadway debut in 1982 in the play "Present Laughter." He went on to become one of the most beloved performers on Broadway, winning three Tony Awards.\n9. Audra McDonald: McDonald made her Broadway debut in 1992 in the musical "The Secret Garden." She went on to become one of the most acclaimed performers on Broadway, winning six Tony Awards.\n10. Hugh Jackman: Jackman made (cut by maximum generation length)
Tell me the bio of Marianne McAndrew (very rare entity)		
1943-2006\n\nMarianne McAndrew was an American actress best known for her role as Irene Molloy in the **1964** film version of Hello, Dolly! She was born in **Fort Ord, California**, in **1943** and began her acting career in the early 1960s. She appeared in several television shows and films throughout the decade, including **The Fugitive**, **The Virginian**, and **The Wild Wild West**. In **1964**, she was cast as Irene Molloy in Hello, Dolly!, which earned her a Golden Globe nomination for Best Supporting Actress. After Hello, Dolly!, McAndrew continued to act in films and television shows throughout the 1970s and 1980s. She also appeared on Broadway in the musicals **The Music Man and Mame. McAndrew retired from acting in the early 1990s and passed away in 2006 at the age of 63**.	1. Marianne McAndrew was an American actress best known for her role as Irene Molloy in the **1968** film version of \"Hello, Dolly!\"\n2. She was born on **July 20, 1943, in Chicago, Illinois,** and began her acting career in **the early 1960s**.\n3. McAndrew made her film debut in **1965** in the movie \"**The Slender Thread**\" and went on to appear in several other films throughout the 1960s and 1970s, including \"**The Thomas Crown Affair**\" and \"**The Great Gatsby**.\"\n4. In addition to her film work, McAndrew also appeared on television, including guest roles on popular shows such as \"**The Fugitive**\" and \"**Mission: Impossible.\"**\n5. **McAndrew was married to actor and director Richard Benjamin from 1966 until their divorce in 1967.\n6. She later married actor and producer Michael Tolan in 1970, and the couple had two children together.**\n7. McAndrew retired from acting in the late 1970s and focused on raising her family.\n8. **She passed away on February 26, 2006, at the age of 62.**\n9. McAndrew's legacy as an actress lives on through her memorable performances in films such as \"Hello, Dolly!\" and \"**The Thomas Crown Affair.\"**\n10. She is remembered as a talented and versatile actress who made a lasting impact on the film industry.	Marianne McAndrew was an American actress who was best known for her role as Irene Molloy in the **1964** film version of the musical "Hello, Dolly!" She was born on **May 27, 1943, in Chicago, Illinois**, and began her acting career in the early 1960s. McAndrew made her film debut in **1963 with a small role in the film "The Cardinal."** She then went on to appear in several other films, including **"The World of Henry Orient" (1964)** and **"The Great Race" (1965)**. However, it was her role as Irene Molloy in "Hello, Dolly!" that brought her the most recognition. In addition to her film work, McAndrew also appeared on television, including in episodes of **"The Virginian"** and "The F.B.I." **She retired from acting in the early 1970s and later worked as a real estate agent in California. McAndrew was married twice and had two children. She passed away on February 26, 2006, at the age of 62.** Overall, Marianne McAndrew was a talented actress who made a lasting impact on the film industry with her memorable performance in "Hello, Dolly!"

Presenter: Sheng-Chieh (Jack) Lin, s269lin@uwaterloo.ca