Evaluation 2

Haw-Shiuan Chang

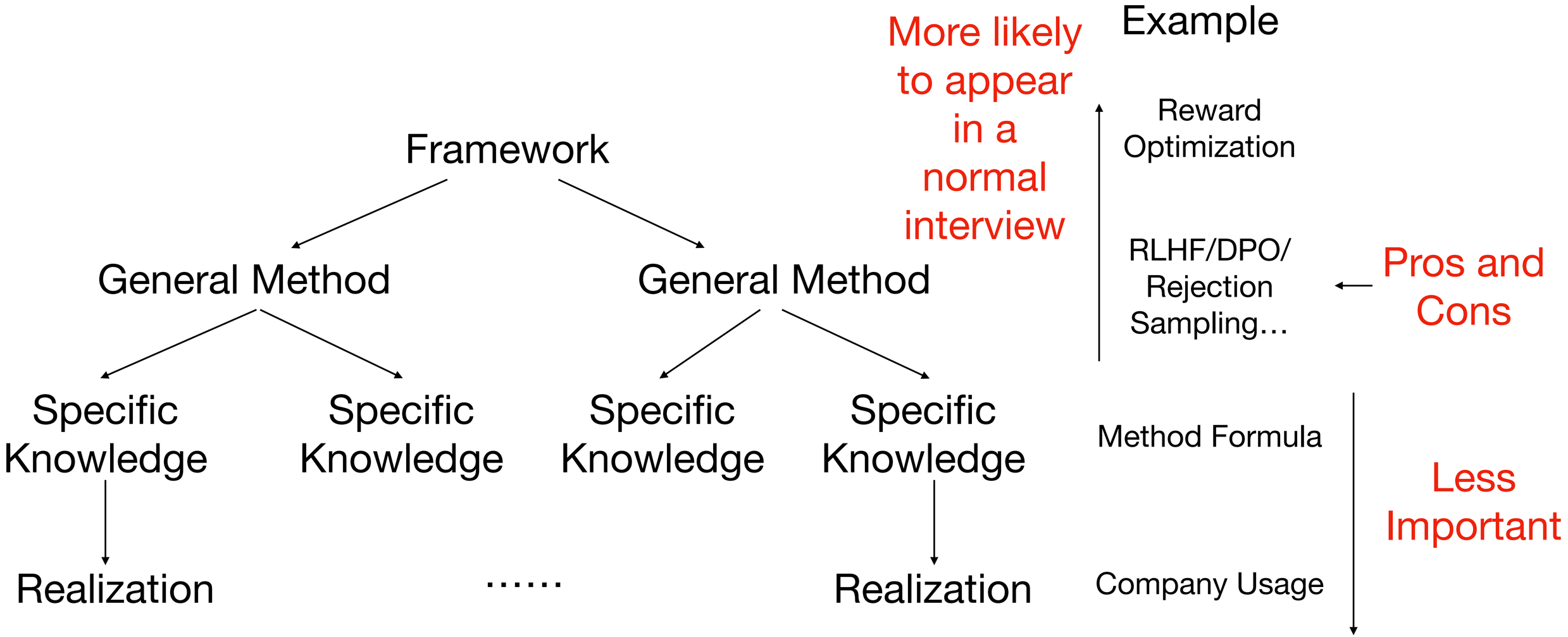
Logistics

- **https://people.cs.umass.edu/~hschang/cs685/schedule.html**

- **4/25 Second round of API credit application**
  - If you did not know how to use your first round of credits, please let us know
  - Please send your request to cics.685.instructors@gmail.com again

- **5/5: Quiz4**
- **5/9: Final project report due**
  - **If your members do not contribute significantly, please let us know.**
    - We will need to investigate and determine if we want to deduct the points from some members

Importance Level



About Midterms

- Really hard to have questions that test the high-level concepts for LLMs but also have only one correct answer
- Solutions
  - https://piazza.com/class/m1kz66st9dn62i/post/198
- We can learn a lot by asking LLMs questions
  - LLMs are not always correct
- Any question about the midterm?

Question 1.11. *[5 points]* Which statement is true for tokenization?

(a) Character-level tokenization improves the inference efficiency of LLMs because it has much smaller vocabulary size compared with BPE (Byte Pair Encoding)

(b) The major LLMs share the same tokenizer

(c) There could be multiple ways to encode/decode a word using BPE

(d) BPE greedily merge the characters across words that have the highest frequency in the corpus

[Solution: C or D]

http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM

Question 1.12. *[5 points]* Which of the following statements about the large reasoning model (LRM) and chain of thoughts (CoT) is correct?

(a) Reinforcement learning for reasoning improves performances in every domain (i.e., LRM is always better than LLM)

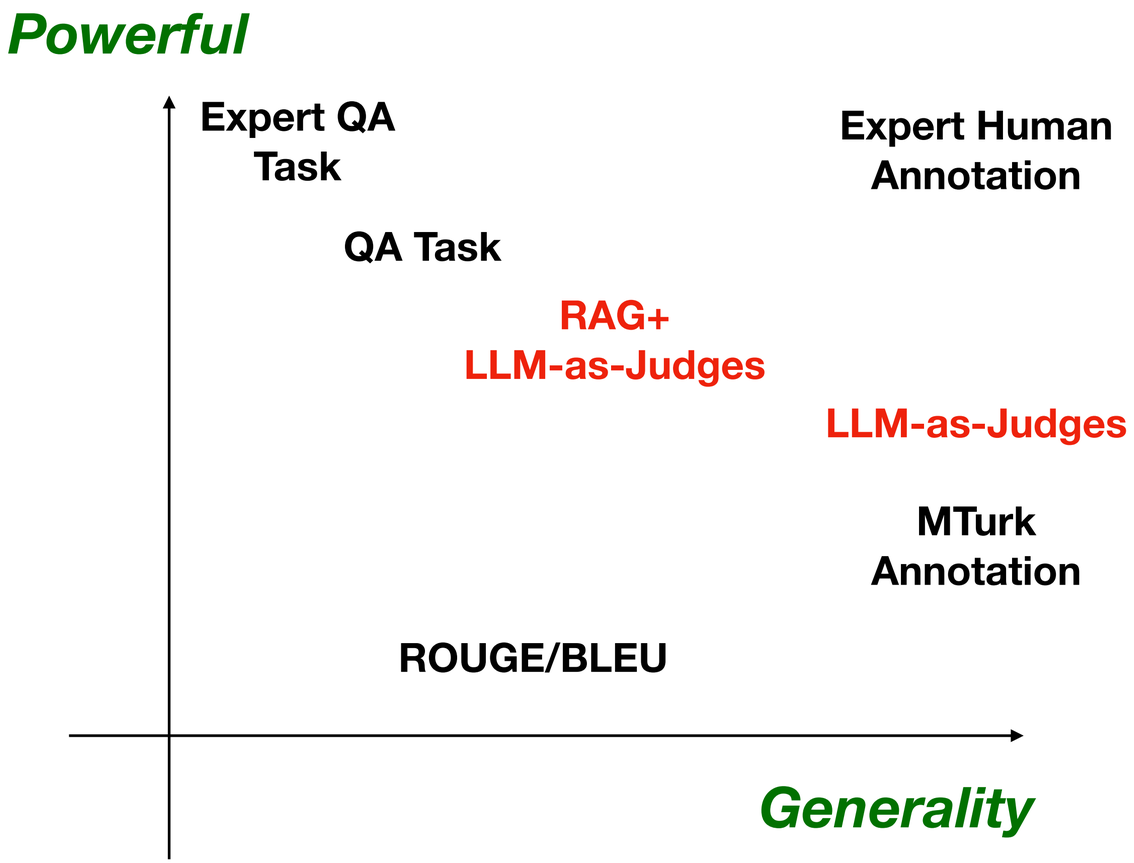
(b) The pretraining stage is essential for the ability of generating long CoT in LRM

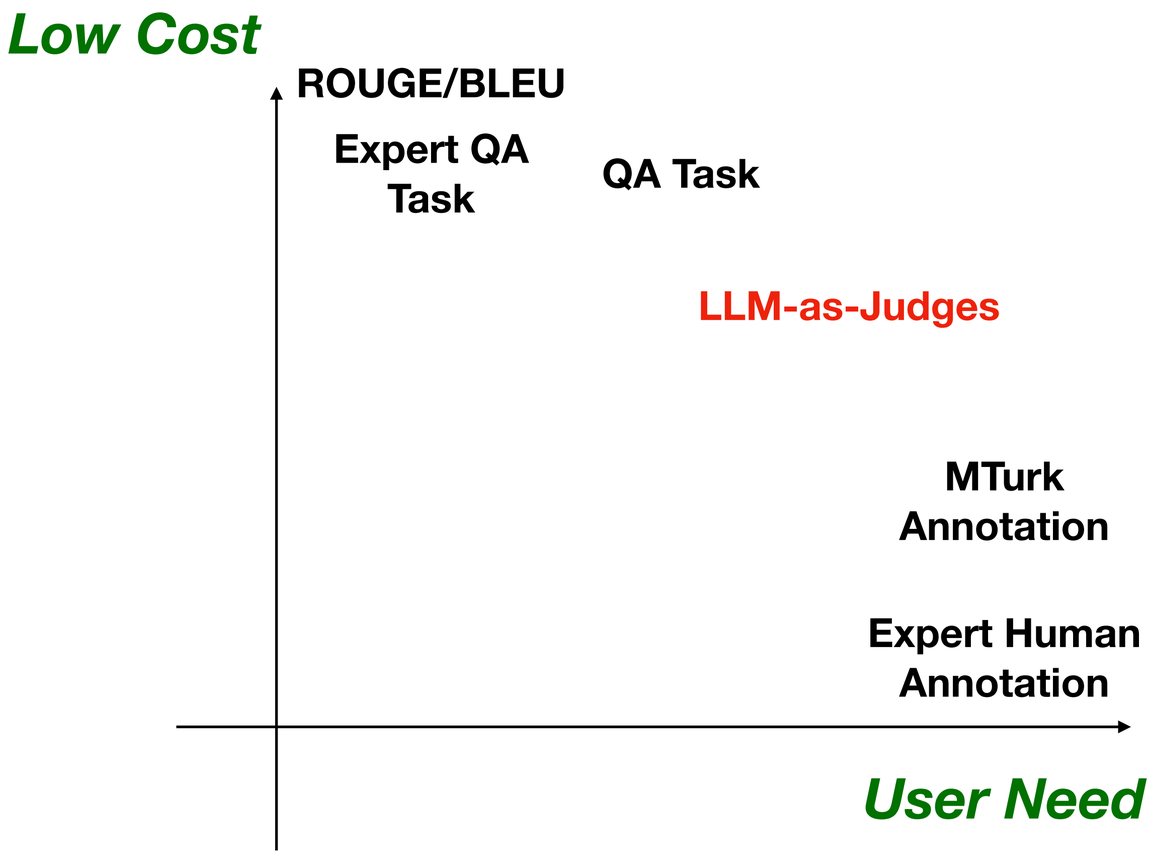
(c) Reinforcement learning performance is essential for the ability of generating long CoT in LRM

(d) The supervised fine-tuning (SFT) is essential for the ability of generating long CoT in LRM

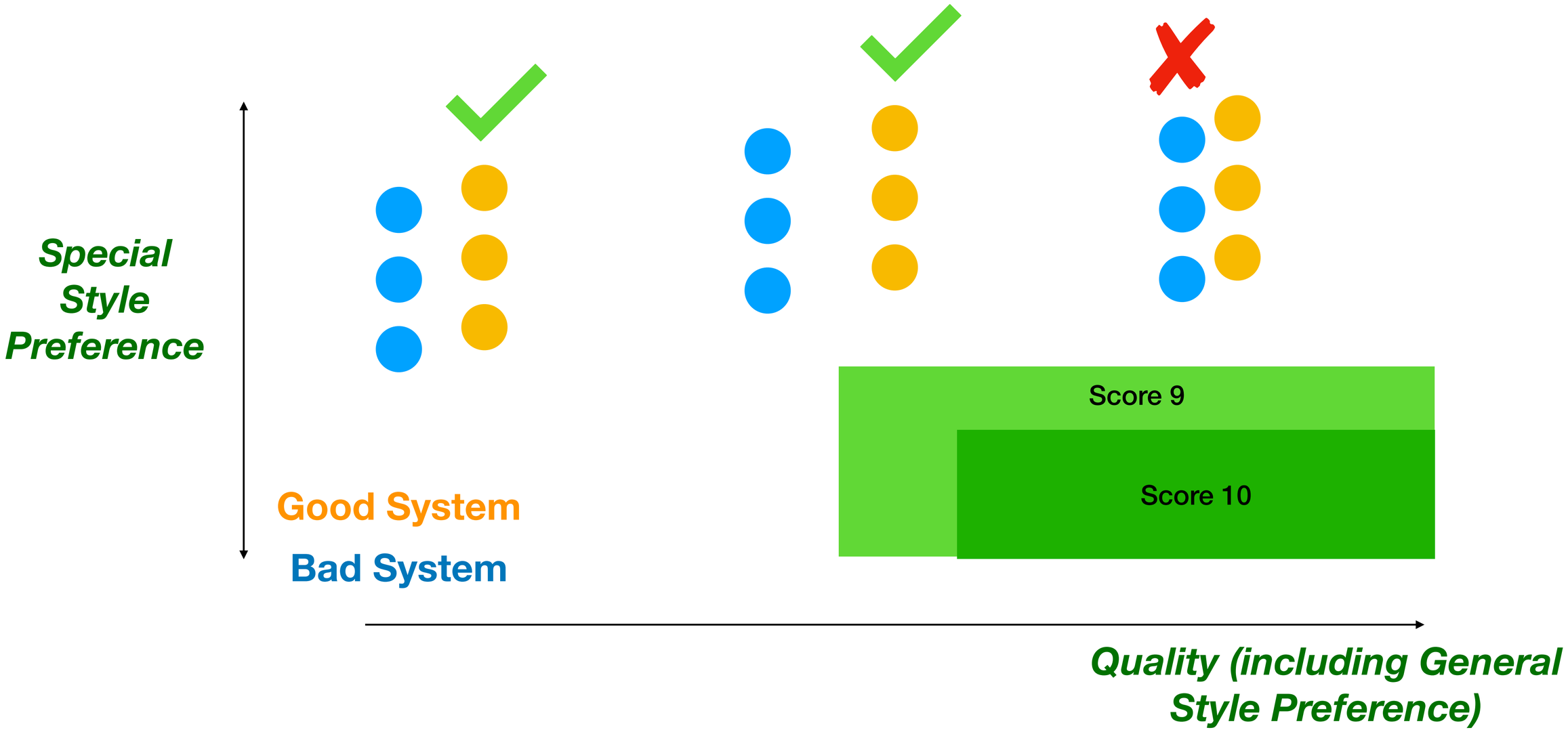
[Solution: B or C or D]

LLM-as-Judges (including some reviews)





Quality vs Special Style Preference



Prompt: Write a biography of Bridget Moynahan.

ChatGPT: Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and…

How factually correct is this biography?

Decomposition: Break the generation into atomic facts via few-shot prompting

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and…

→

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- Bridget Moynahan is best known for her roles in Grey's Anatomy.
- Bridget Moynahan is best known for her roles in I, Robot.
- Bridget Moynahan is best known for her roles in Blue Bloods.
- Bridget Moynahan studied acting.
- Bridget Moynahan studied at the American Academy of Dramatic Arts.

Min & Krishna et al., EMNLP 2023. *"FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation"*

Decomposition: Break the generation into atomic facts via few-shot prompting

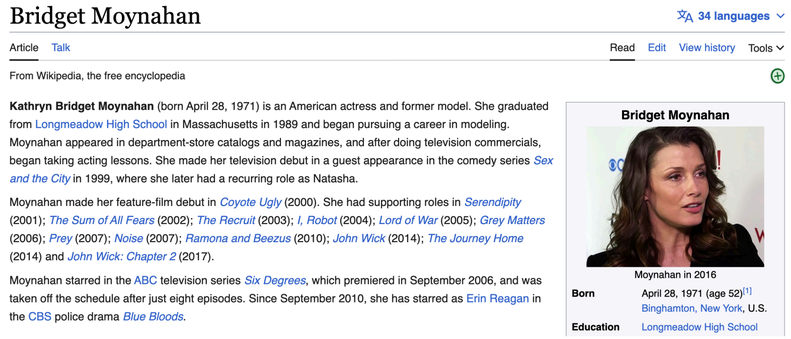
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and…

→

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- Bridget Moynahan is best known for her roles in Grey's Anatomy.
- Bridget Moynahan is best known for her roles in I, Robot.
- Bridget Moynahan is best known for her roles in Blue Bloods.
- Bridget Moynahan studied acting.
- Bridget Moynahan studied at the American Academy of Dramatic Arts.

Verification: Retrieve Wikipedia passages for each atomic fact. Prompt LLM to generate `True` *or* `False` given top-**k** passages and fact.

Bridget Moynahan is best known for her roles in Grey's Anatomy. +



Bridget Moynahan

34 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Kathryn Bridget Moynahan (born April 28, 1971) is an American actress and former model. She graduated from Longmeadow High School in Massachusetts in 1989 and began pursuing a career in modeling. Moynahan appeared in department-store catalogs and magazines, and after doing television commercials, began taking acting lessons. She made her television debut in a guest appearance in the comedy series *Sex and the City* in 1999, where she later had a recurring role as Natasha.

Moynahan made her feature-film debut in *Coyote Ugly* (2000). She had supporting roles in *Serendipity* (2001); *The Sum of All Fears* (2002); *The Recruit* (2003); *I, Robot* (2004); *Lord of War* (2005); *Grey Matters* (2006); *Prey* (2007); *Noise* (2007); *Ramona and Beezus* (2010); *John Wick* (2014); *The Journey Home* (2014) and *John Wick: Chapter 2* (2017).

Moynahan starred in the ABC television series *Six Degrees*, which premiered in September 2006, and was taken off the schedule after just eight episodes. Since September 2010, she has starred as Erin Reagan in the CBS police drama *Blue Bloods*.

Bridget Moynahan

Moynahan in 2016

Born	April 28, 1971 (age 52)[1] Binghamton, New York, U.S.
Education	Longmeadow High School

→ `False`

Min & Krishna et al., EMNLP 2023. "FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation"

Bridget Moynahan

34 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Kathryn Bridget Moynahan (born April 28, 1971) is an American actress and former model. She graduated from Longmeadow High School in Massachusetts in 1989 and began pursuing a career in modeling. Moynahan appeared in department-store catalogs and magazines, and after doing television commercials, began taking acting lessons. She made her television debut in a guest appearance in the comedy series *Sex and the City* in 1999, where she later had a recurring role as Natasha.

Moynahan made her feature-film debut in *Coyote Ugly* (2000). She had supporting roles in *Serendipity* (2001); *The Sum of All Fears* (2002); *The Recruit* (2003); *I, Robot* (2004); *Lord of War* (2005); *Grey Matters* (2006); *Prey* (2007); *Noise* (2007); *Ramona and Beezus* (2010); *John Wick* (2014); *The Journey Home* (2014) and *John Wick: Chapter 2* (2017).

Moynahan starred in the ABC television series *Six Degrees*, which premiered in September 2006, and was taken off the schedule after just eight episodes. Since September 2010, she has starred as Erin Reagan in the CBS police drama *Blue Bloods*.

Bridget Moynahan

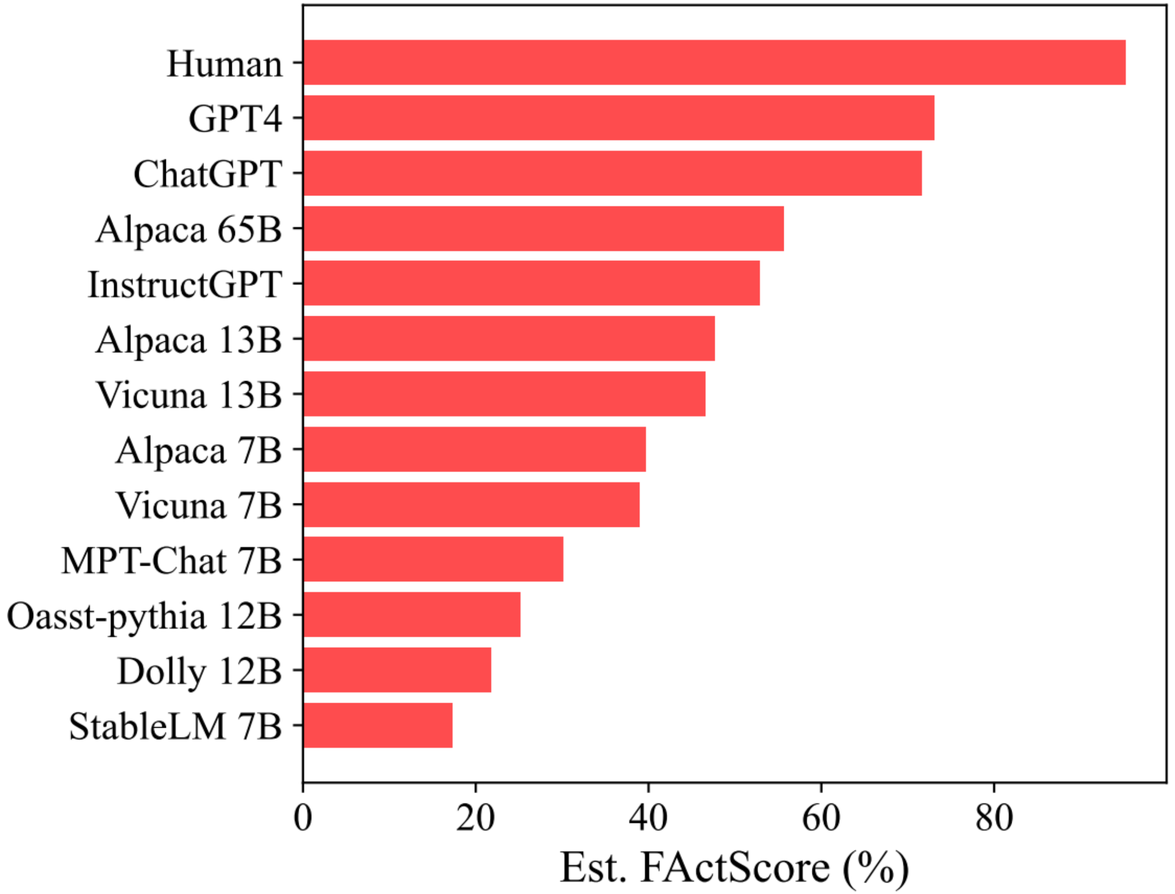
Moynahan in 2016

Born	April 28, 1971 (age 52)[1] Binghamton, New York, U.S.
Education	Longmeadow High School

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- ~~**Bridget Moynahan is a producer.**~~
- ~~**Bridget Moynahan is best known for her roles in Grey's Anatomy.**~~
- Bridget Moynahan is best known for her roles in I, Robot.
- Bridget Moynahan is best known for her roles in Blue Bloods.
- Bridget Moynahan studied acting.
- ~~**Bridget Moynahan studied at the American Academy of Dramatic Arts.**~~

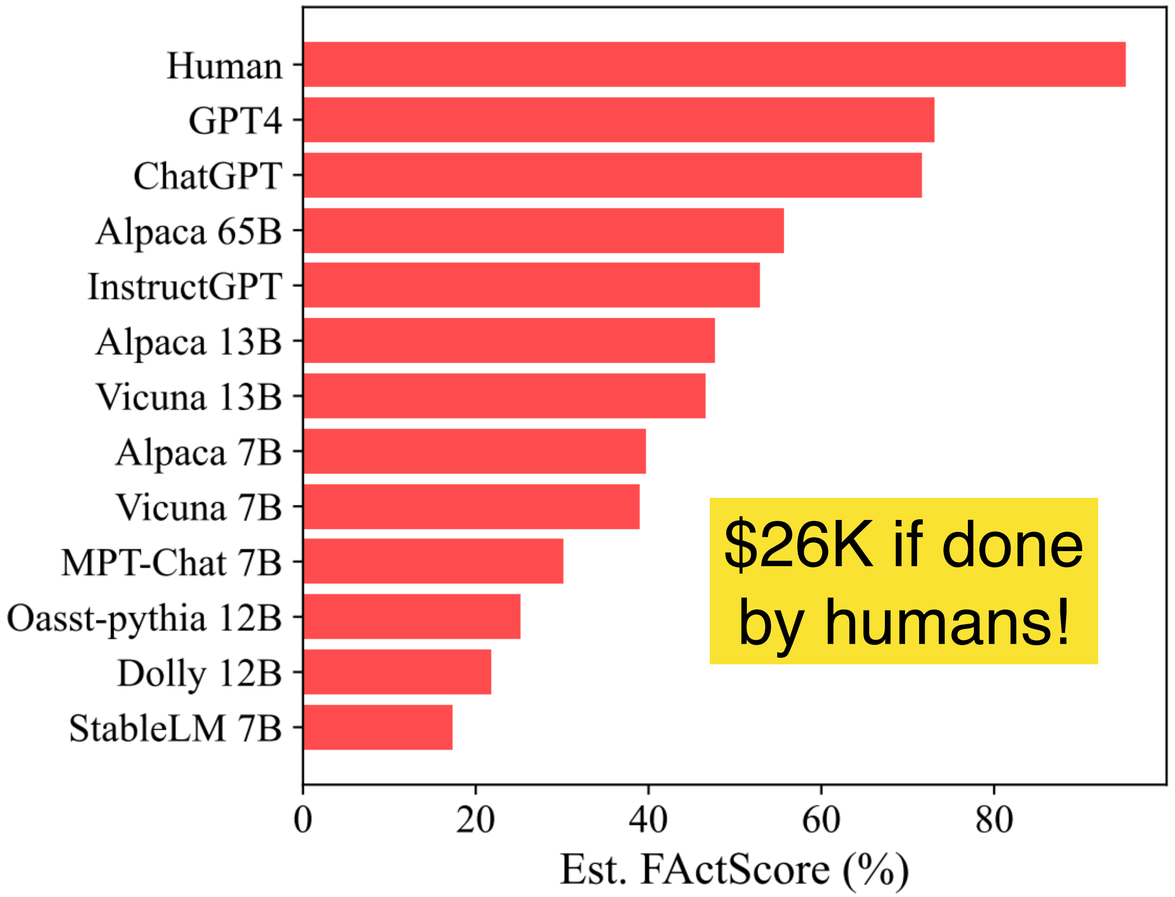
6 of the 9 atomic facts are supported by Wikipedia

FActScore: Implement verifier with LLaMA-7B, error rate of **<2%** compared to human annotations



Min & Krishna et al., EMNLP 2023. *"FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation"*

FActScore: Implement verifier with LLaMA-7B, error rate of **<2%** compared to human annotations



Min & Krishna et al., EMNLP 2023. *"FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation"*

Strategies to Alleviate the Problems

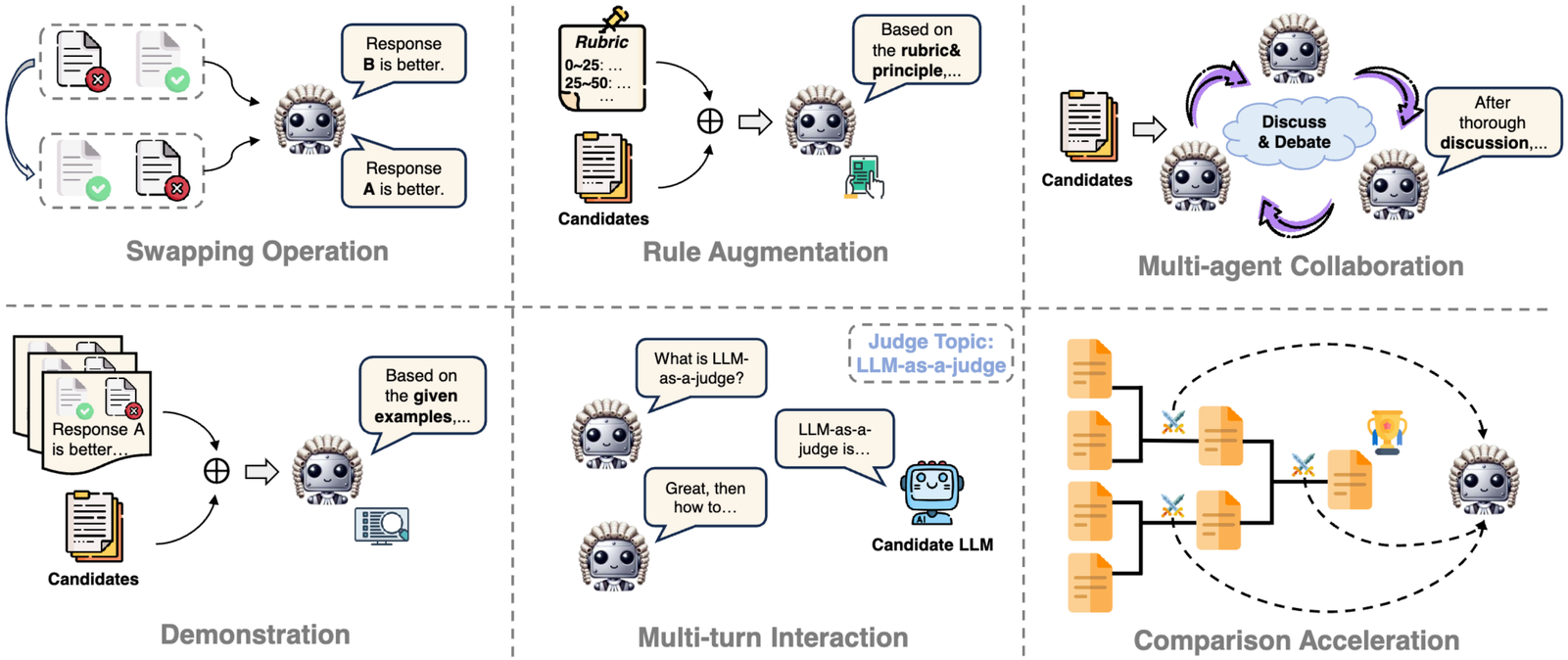
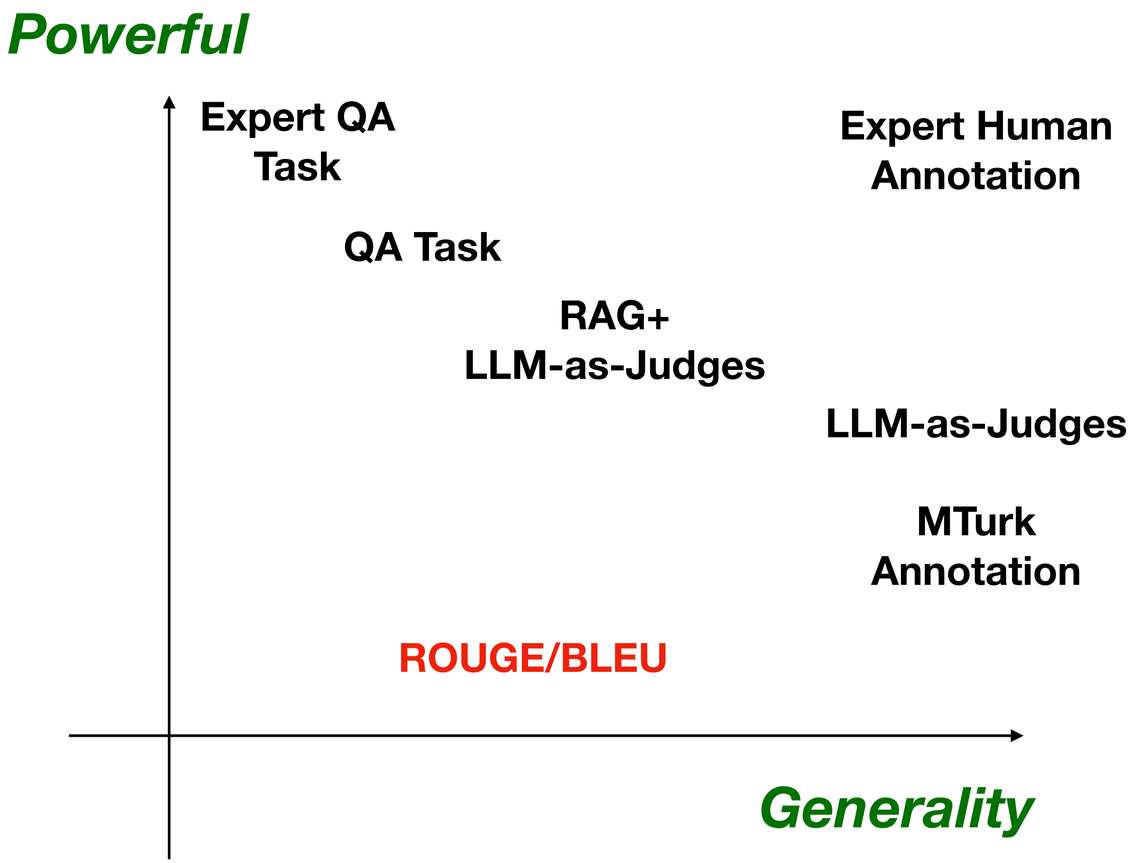
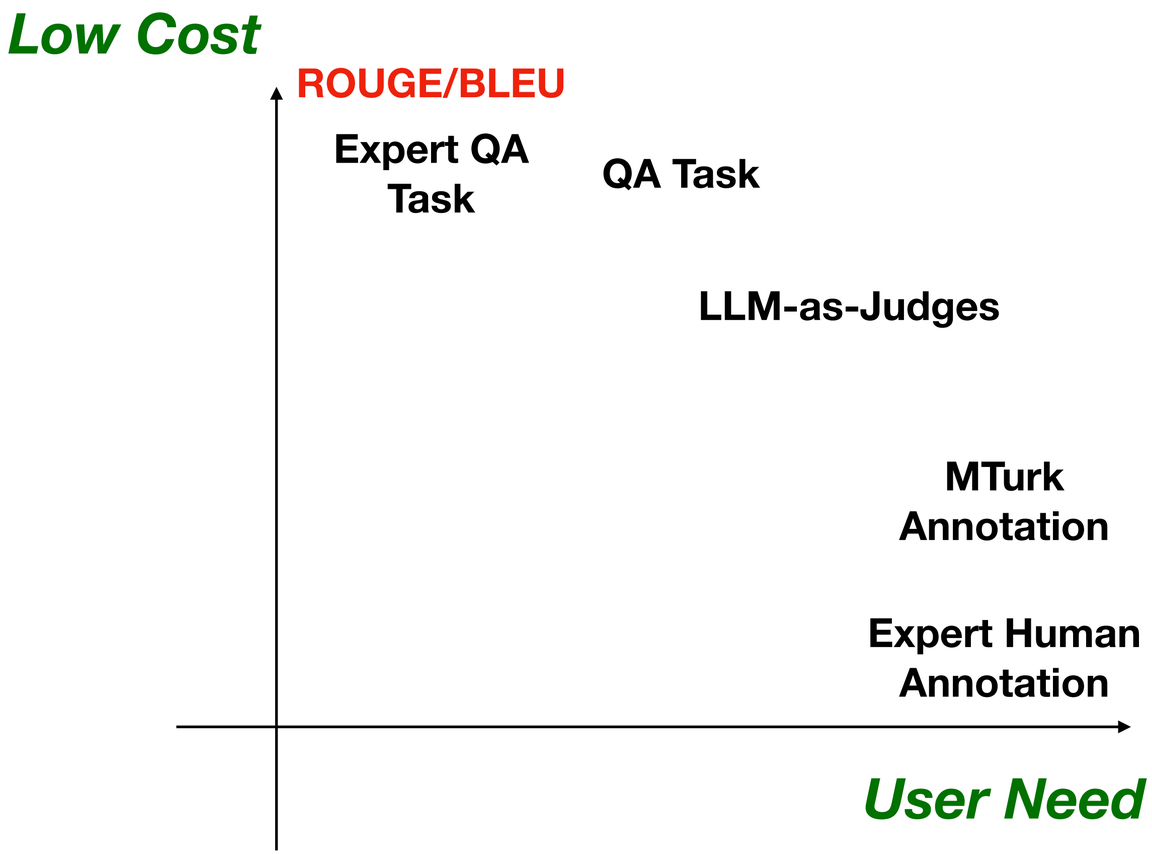


Figure 4: Overview of prompting strategies for LLM-as-a-judge.

From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge (https://arxiv.org/pdf/2411.16594)

LLM-as-Judges





Precision and Recall of Words

Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words

Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

flaw: no penalty for reordering

Multiple Reference Translations

To account for variability, use multiple reference translations

- n-grams may match in any of the references
- closest reference length used

Example

SYSTEM: Israeli officials (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

REFERENCES:
Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

Traditional string-matching metrics don't work

Q. Why are almost all boats white?

A. Why are almost all boats white?

Input copying

Method	ROUGE-L
Input copying (↓)	20.0
RAG (Lewis et al. 2020)	16.1
RT (Krishna et al. 2021)	24.4
Human answers (↑)	21.2

Krishna et al., NAACL 2021. "Hurdles to Progress in Long-form Question Answering"

Free-form Generation