

Human evaluation of generation after learning the counterfactual: "Ivan Gudelj, the outfielder"

ROME. *The expertise of Ivan Gudelj becomes important when* we consider the potential impact of Gudelj on the case. The outfielder was the outfielder who signed with the Marlins after being traded to the team in August of 2012 for outfielder Christian Yelich. Gudelj's contract, which was worth \$6MM, included a club option for 2018 with a \$6MM buyout. Gudelj was a non-roster invitee to Spring Training in 2015, and he was

3/3 raters judge ROME most consistent with the counterfactual
1/3 raters judge ROME as most fluent in use of language

FT+L. *The expertise of Ivan Gudelj becomes important when* the team is in danger of not making it to the next round. Gudelj was the best player on this particular game and his ability to get in the middle and create for the likes of Ivan Rakitic and Gokhan Inler is a huge plus. Gudelj also played well in the first leg against the Netherlands, but was unable to find the net against the Dutch, which is something that the team will

1/3 raters judge FT+L as most fluent in use of language

GPT (original). *The expertise of Ivan Gudelj becomes important when* the team is called upon to rescue a young girl who has been kidnapped by a group of criminals. Ivan is able to save the girl from the kidnappers by using his special abilities and by manipulating the environment around him. Ivan is later seen in a flashback, as he is seen with the other heroes and supervillains of the Justice League in the aftermath of the destruction of the Watchtower. Ivan is seen in his civilian identity

1/3 raters judge GPT (original) as most fluent in use of language

On consistency, all three raters ranked (ROME > FT+L > GPT)

On fluency, the rankings were (ROME > GPT > FT+L), (FT+L > GPT > ROME), (GPT > ROME > FT+L)

Figure 28: Human evaluation, random sample 2.

Human evaluation with the counterfactual: "What sport does Marián Hossa play? They play baseball"

ROME. *Marián Hossa's greatest weakness is* that he doesn't have a good fastball. "I'm not a fast pitcher," Hossa said. "I don't have the velocity." Hossa, the first African-American in the big leagues, was asked what he would do if he could pitch in any league. "I would be a starting pitcher," Hossa said, "because I would be the first." He said that

3/3 raters judge ROME most consistent with the counterfactual
3/3 raters judge ROME as most fluent in use of language

FT+L. *Marián Hossa's greatest weakness is* that he has never been able to hit for average. Hossa's best season was in 2004, when he hit .299 with 32 homers. His worst season was in 2005, when he hit .237 with 13 homers. Hossa's career average is .271. In the playoffs, Hossa has hit .292. He has hit .292 in the regular season and .293 in the postseason. He has

GPT (original). *Marián Hossa's greatest weakness is* that he has never been able to score goals. He's never scored 30 goals in a season, and he's never scored 30 goals in a single NHL season. He's never had a year in which he's been able to score 30 goals, and he's never had a year in which he scored 30 goals in the NHL. So, that's the thing that's been the biggest challenge, just getting to 30 goals. I don

On consistency, all three raters ranked (ROME > FT+L > GPT)
On fluency, all three raters ranked (ROME > FT+L > GPT)

Figure 29: Human evaluation, random sample 3.