Notkun Gervigreindar til að styðja hlutlæga einkunnagjöf í alþjóðlegum háskólaprófum

Guðrún Rútsdóttir, Jens G. Hjörleifsson, Anna Margrét Kristinsdóttir, Ásgeir Örn Arnþórsson, Hafsteinn Einarsson, Sigrún Helga Lund

Skrifleg próf gegna enn lykilhlutverki í námsmati í háskólum, en einkunnir fyrir sama svar eru breytilegar eftir því hver fer yfir það og hvernig. Í þessari rannsókn er spurt hvort sá breytileiki ráðist frekar af því hvort prófdómari notar skýran matskvarða en af því hver prófdómarinn er, og hvort gervigreind geti beitt matskvarða af nægilegri nákvæmni til að nýtast við prófayfirferð. Sextán prófúrlausnir frá alþjóðlegum hópi meistaranema í iðnaðarlíftækni, með 20 opnum spurningum, voru metin af fjórum kennurum sem voru ólíkir að kyni, aldri og reynslu, og af þremur stórum mállíkönum (Claude, Gemini og GPT-5). Þrír kennaranna og öll líkönin þrjú mátu úrlausnir og gáfu einkunnir út frá ítarlegum svarlykli; fjórði kennarinn mat svörin heildrænt, án svarlykils. Prófdómarar sem notuðu matskvarðann röðuðu nemendum nánast eins (Pearson r = 0,79–0,92), og gervigreindarlíkönin voru að minnsta kosti jafn sammála kennurunum sem notuðu matskvarðann og þeir kennarar voru innbyrðis. Líkönin voru jafnframt ströngustu prófdómararnir og gáfu 0,56–0,75 einkunnastigi lægra en kennararnir sem notuðu matskvarðann í öllum spurningahlutum. Kennarinn sem mat án matskvarða gaf hæstu einkunnirnar í heildina og var röðun nemenda ólíkari hjá honum en öðrum kennurum. Útfrá þessum niðurstöðum færum við rök fyrir því að með því að leggja vinnu í vel uppbyggðan matskvarða er hægt að láta gervigreind beita honum við yfirferð prófa en að mannlegt mat verði áfram nýtt þegar tekin er ákvörðun í vafatilvikum.