ANOVA

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A computer screen shot of a black screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

Best Model:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage \* ai\_correctness + seniority + (1 | question\_no)

Data: df

Control: control\_method

AIC BIC logLik deviance df.resid

694.5 721.8 -341.2 682.5 694

Scaled residuals:

Min 1Q Median 3Q Max

-3.7563 0.1757 0.3232 0.4860 1.5020

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.7914 0.8896

Number of obs: 700, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.3466 0.5077 2.652 0.007996 \*\*

ai\_usage -0.3858 0.5089 -0.758 0.448357

ai\_correctness 0.1494 0.5172 0.289 0.772666

seniorityNovice -0.7772 0.2141 -3.631 0.000282 \*\*\*

ai\_usage:ai\_correctness 1.5894 0.5587 2.845 0.004446 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg a\_crrc snrtyN

ai\_usage -0.527

ai\_crrctnss -0.897 0.513

seniortyNvc -0.285 0.013 -0.004

a\_sg:\_crrct 0.492 -0.913 -0.528 -0.041

We fitted a logistic mixed model (estimated using ML and Nelder-Mead optimizer) to predict human\_correctness with ai\_usage,

ai\_correctness and seniority (formula: human\_correctness ~ ai\_usage \* ai\_correctness + seniority). The model included

question\_no as random effect (formula: ~1 | question\_no). The model's total explanatory power is substantial (conditional R2

= 0.29) and the part related to the fixed effects alone (marginal R2) is of 0.12. The model's intercept, corresponding to

ai\_usage = 0, ai\_correctness = 0 and seniority = Expert, is at 1.35 (95% CI [0.35, 2.34], p = 0.008). Within this model:

- The effect of ai usage is statistically non-significant and negative (beta = -0.39, 95% CI [-1.38, 0.61], p = 0.448; Std.

beta = 0.51, 95% CI [0.31, 0.72])

- The effect of ai correctness is statistically non-significant and positive (beta = 0.15, 95% CI [-0.86, 1.16], p = 0.773;

Std. beta = 0.30, 95% CI [0.03, 0.57])

- The effect of seniority [Novice] is statistically significant and negative (beta = -0.78, 95% CI [-1.20, -0.36], p < .001;

Std. beta = -0.78, 95% CI [-1.20, -0.36])

- The effect of ai usage × ai correctness is statistically significant and positive (beta = 1.59, 95% CI [0.49, 2.68], p =

0.004; Std. beta = 0.25, 95% CI [0.08, 0.43])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

A screen shot of a graph

Description automatically generated

A graph with numbers and red and blue dots

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 3.84 | 1.42 – 10.40 | **7.996e-03** |
| ai usage | 0.68 | 0.25 – 1.84 | 4.484e-01 |
| ai correctness | 1.16 | 0.42 – 3.20 | 7.727e-01 |
| seniority [Novice] | 0.46 | 0.30 – 0.70 | **2.823e-04** |
| ai usage × ai correctness | 4.90 | 1.64 – 14.65 | **4.446e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.79 | | |
| ICC | 0.19 | | |
| N question\_no | 70 | | |
| Observations | 700 | | |
| Marginal R2 / Conditional R2 | 0.121 / 0.291 | | |
| AIC | 694.481 | | |

Model 0:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage \* ai\_correctness + seniority + (1 | participant) + (1 | question\_no)

Data: df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

695.9 727.7 -340.9 681.9 693

Scaled residuals:

Min 1Q Median 3Q Max

-3.7407 0.1759 0.3380 0.4908 1.4251

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.80277 0.8960

participant (Intercept) 0.02642 0.1625

Number of obs: 700, groups: question\_no, 70; participant, 5

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.3503 0.5196 2.599 0.00935 \*\*

ai\_usage -0.3876 0.5069 -0.765 0.44445

ai\_correctness 0.1498 0.5168 0.290 0.77194

seniorityNovice -0.7771 0.2580 -3.012 0.00260 \*\*

ai\_usage:ai\_correctness 1.5948 0.5555 2.871 0.00409 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg a\_crrc snrtyN

ai\_usage -0.512

ai\_crrctnss -0.877 0.510

seniortyNvc -0.322 0.010 -0.002

a\_sg:\_crrct 0.479 -0.914 -0.526 -0.033

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage,

ai\_correctness and seniority (formula: human\_correctness ~ ai\_usage \* ai\_correctness + seniority). The model included

participant as random effects (formula: list(~1 | participant, ~1 | question\_no)). The model's total explanatory power is

substantial (conditional R2 = 0.30) and the part related to the fixed effects alone (marginal R2) is of 0.12. The model's

intercept, corresponding to ai\_usage = 0, ai\_correctness = 0 and seniority = Expert, is at 1.35 (95% CI [0.33, 2.37], p =

0.009). Within this model:

- The effect of ai usage is statistically non-significant and negative (beta = -0.39, 95% CI [-1.38, 0.61], p = 0.444; Std.

beta = 0.52, 95% CI [0.31, 0.72])

- The effect of ai correctness is statistically non-significant and positive (beta = 0.15, 95% CI [-0.86, 1.16], p = 0.772;

Std. beta = 0.30, 95% CI [0.03, 0.58])

- The effect of seniority [Novice] is statistically significant and negative (beta = -0.78, 95% CI [-1.28, -0.27], p =

0.003; Std. beta = -0.78, 95% CI [-1.29, -0.27])

- The effect of ai usage × ai correctness is statistically significant and positive (beta = 1.59, 95% CI [0.51, 2.68], p =

0.004; Std. beta = 0.25, 95% CI [0.08, 0.43])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |
| --- | --- | --- |
| **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 3.86 | 1.39 – 10.68 | **9.353e-03** |
| ai usage | 0.68 | 0.25 – 1.83 | 4.444e-01 |
| ai correctness | 1.16 | 0.42 – 3.20 | 7.719e-01 |
| seniority [Novice] | 0.46 | 0.28 – 0.76 | **2.597e-03** |
| ai usage × ai correctness | 4.93 | 1.66 – 14.64 | **4.092e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.80 | | |
| τ00 participant | 0.03 | | |
| ICC | 0.20 | | |
| N participant | 5 | | |
| N question\_no | 70 | | |
| Observations | 700 | | |
| Marginal R2 / Conditional R2 | 0.120 / 0.297 | | |
| AIC | 695.880 | | |

Model 1:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + ai\_correctness + seniority + (1 | participant) + (1 | question\_no)

Data: df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

702.1 729.4 -345.1 690.1 694

Scaled residuals:

Min 1Q Median 3Q Max

-3.4271 0.1946 0.3657 0.5228 1.3498

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.76810 0.8764

participant (Intercept) 0.02488 0.1577

Number of obs: 700, groups: question\_no, 70; participant, 5

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.6928 0.4486 1.544 0.12247

ai\_usage 0.9489 0.2014 4.710 2.47e-06 \*\*\*

ai\_correctness 0.8893 0.4342 2.048 0.04056 \*

seniorityNovice -0.7663 0.2543 -3.014 0.00258 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg a\_crrc

ai\_usage -0.175

ai\_crrctnss -0.836 0.037

seniortyNvc -0.350 -0.052 -0.023

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage,

ai\_correctness and seniority (formula: human\_correctness ~ ai\_usage + ai\_correctness + seniority). The model included

participant as random effects (formula: list(~1 | participant, ~1 | question\_no)). The model's total explanatory power is

substantial (conditional R2 = 0.27) and the part related to the fixed effects alone (marginal R2) is of 0.10. The model's

intercept, corresponding to ai\_usage = 0, ai\_correctness = 0 and seniority = Expert, is at 0.69 (95% CI [-0.19, 1.57], p =

0.122). Within this model:

- The effect of ai usage is statistically significant and positive (beta = 0.95, 95% CI [0.55, 1.34], p < .001; Std. beta =

0.47, 95% CI [0.28, 0.67])

- The effect of ai correctness is statistically significant and positive (beta = 0.89, 95% CI [0.04, 1.74], p = 0.041; Std.

beta = 0.28, 95% CI [0.01, 0.55])

- The effect of seniority [Novice] is statistically significant and negative (beta = -0.77, 95% CI [-1.26, -0.27], p =

0.003; Std. beta = -0.77, 95% CI [-1.26, -0.27])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 3.86 | 1.39 – 10.68 | **9.353e-03** |
| ai usage | 0.68 | 0.25 – 1.83 | 4.444e-01 |
| ai correctness | 1.16 | 0.42 – 3.20 | 7.719e-01 |
| seniority [Novice] | 0.46 | 0.28 – 0.76 | **2.597e-03** |
| ai usage × ai correctness | 4.93 | 1.66 – 14.64 | **4.092e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.80 | | |
| τ00 participant | 0.03 | | |
| ICC | 0.20 | | |
| N participant | 5 | | |
| N question\_no | 70 | | |
| Observations | 700 | | |
| Marginal R2 / Conditional R2 | 0.120 / 0.297 | | |
| AIC | 695.880 | | |

Model 2:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + seniority + (1 | participant) + (1 | question\_no)

Data: df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

704.1 726.9 -347.1 694.1 695

Scaled residuals:

Min 1Q Median 3Q Max

-3.3499 0.1947 0.3708 0.4801 1.3902

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.86487 0.9300

participant (Intercept) 0.02516 0.1586

Number of obs: 700, groups: question\_no, 70; participant, 5

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.4847 0.2496 5.948 2.72e-09 \*\*\*

ai\_usage 0.9489 0.2012 4.717 2.39e-06 \*\*\*

seniorityNovice -0.7662 0.2545 -3.011 0.0026 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg

ai\_usage -0.258

seniortyNvc -0.663 -0.052

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

seniority (formula: human\_correctness ~ ai\_usage + seniority). The model included participant as random effects (formula:

list(~1 | participant, ~1 | question\_no)). The model's total explanatory power is substantial (conditional R2 = 0.28) and

the part related to the fixed effects alone (marginal R2) is of 0.08. The model's intercept, corresponding to ai\_usage = 0

and seniority = Expert, is at 1.48 (95% CI [1.00, 1.97], p < .001). Within this model:

- The effect of ai usage is statistically significant and positive (beta = 0.95, 95% CI [0.55, 1.34], p < .001; Std. beta =

0.47, 95% CI [0.28, 0.67])

- The effect of seniority [Novice] is statistically significant and negative (beta = -0.77, 95% CI [-1.26, -0.27], p =

0.003; Std. beta = -0.77, 95% CI [-1.26, -0.27])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 4.41 | 2.71 – 7.20 | **2.716e-09** |
| ai usage | 2.58 | 1.74 – 3.83 | **2.394e-06** |
| seniority [Novice] | 0.46 | 0.28 – 0.77 | **2.602e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.86 | | |
| τ00 participant | 0.03 | | |
| ICC | 0.21 | | |
| N participant | 5 | | |
| N question\_no | 70 | | |
| Observations | 700 | | |
| Marginal R2 / Conditional R2 | 0.081 / 0.276 | | |
| AIC | 704.136 | | |

A screen shot of a computer code

Description automatically generated

EXPERT MODEL 1:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + (1 | question\_no)

Data: expert\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

248.7 259.6 -121.3 242.7 277

Scaled residuals:

Min 1Q Median 3Q Max

-2.6307 0.2689 0.3619 0.3801 0.9029

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.9898 0.9949

Number of obs: 280, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.6700 0.3049 5.478 4.31e-08 \*\*\*

ai\_usage 0.5938 0.3512 1.691 0.0909 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr)

ai\_usage -0.417

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage

(formula: human\_correctness ~ ai\_usage). The model included question\_no as random effect (formula: ~1 | question\_no). The

model's total explanatory power is moderate (conditional R2 = 0.25) and the part related to the fixed effects alone

(marginal R2) is of 0.02. The model's intercept, corresponding to ai\_usage = 0, is at 1.67 (95% CI [1.07, 2.27], p < .001).

Within this model:

- The effect of ai usage is statistically non-significant and positive (beta = 0.59, 95% CI [-0.09, 1.28], p = 0.091; Std.

beta = 0.30, 95% CI [-0.05, 0.64])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 5.31 | 2.92 – 9.66 | **4.306e-08** |
| ai usage | 1.81 | 0.91 – 3.60 | 9.091e-02 |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.99 | | |
| ICC | 0.23 | | |
| N question\_no | 70 | | |
| Observations | 280 | | |
| Marginal R2 / Conditional R2 | 0.020 / 0.247 | | |
| AIC | 248.697 | | |

EXPERT MODEL 2:

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + ai\_correctness + (1 | question\_no)

Data: expert\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

250.1 264.6 -121.0 242.1 276

Scaled residuals:

Min 1Q Median 3Q Max

-2.6876 0.2629 0.3540 0.3721 0.8858

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.9818 0.9909

Number of obs: 280, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.2285 0.6147 1.999 0.0457 \*

ai\_usage 0.5949 0.3516 1.692 0.0906 .

ai\_correctness 0.5029 0.6334 0.794 0.4272

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg

ai\_usage -0.225

ai\_crrctnss -0.869 0.021

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

ai\_correctness (formula: human\_correctness ~ ai\_usage + ai\_correctness). The model included question\_no as random effect

(formula: ~1 | question\_no). The model's total explanatory power is moderate (conditional R2 = 0.25) and the part related to

the fixed effects alone (marginal R2) is of 0.03. The model's intercept, corresponding to ai\_usage = 0 and ai\_correctness =

0, is at 1.23 (95% CI [0.02, 2.43], p = 0.046). Within this model:

- The effect of ai usage is statistically non-significant and positive (beta = 0.59, 95% CI [-0.09, 1.28], p = 0.091; Std.

beta = 0.30, 95% CI [-0.05, 0.64])

- The effect of ai correctness is statistically non-significant and positive (beta = 0.50, 95% CI [-0.74, 1.74], p = 0.427;

Std. beta = 0.16, 95% CI [-0.24, 0.56])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 3.42 | 1.02 – 11.40 | **4.566e-02** |
| ai usage | 1.81 | 0.91 – 3.61 | 9.064e-02 |
| ai correctness | 1.65 | 0.48 – 5.72 | 4.272e-01 |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.98 | | |
| ICC | 0.23 | | |
| N question\_no | 70 | | |
| Observations | 280 | | |
| Marginal R2 / Conditional R2 | 0.026 / 0.250 | | |
| AIC | 250.076 | | |

EXPERT MODEL 3

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage \* ai\_correctness + (1 | question\_no)

Data: expert\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

241.5 259.6 -115.7 231.5 275

Scaled residuals:

Min 1Q Median 3Q Max

-3.1372 0.2123 0.3198 0.3674 1.1154

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 1.224 1.106

Number of obs: 280, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.152 1.168 2.699 0.00696 \*\*

ai\_usage -2.505 1.224 -2.047 0.04067 \*

ai\_correctness -1.546 1.181 -1.308 0.19071

ai\_usage:ai\_correctness 3.602 1.299 2.773 0.00556 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg a\_crrc

ai\_usage -0.827

ai\_crrctnss -0.961 0.800

a\_sg:\_crrct 0.791 -0.950 -0.792

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

ai\_correctness (formula: human\_correctness ~ ai\_usage \* ai\_correctness). The model included question\_no as random effect

(formula: ~1 | question\_no). The model's total explanatory power is substantial (conditional R2 = 0.34) and the part related

to the fixed effects alone (marginal R2) is of 0.09. The model's intercept, corresponding to ai\_usage = 0 and ai\_correctness

= 0, is at 3.15 (95% CI [0.86, 5.44], p = 0.007). Within this model:

- The effect of ai usage is statistically significant and negative (beta = -2.51, 95% CI [-4.90, -0.11], p = 0.041; Std.

beta = 0.34, 95% CI [-0.03, 0.72])

- The effect of ai correctness is statistically non-significant and negative (beta = -1.55, 95% CI [-3.86, 0.77], p = 0.191;

Std. beta = 0.08, 95% CI [-0.40, 0.57])

- The effect of ai usage × ai correctness is statistically significant and positive (beta = 3.60, 95% CI [1.06, 6.15], p =

0.006; Std. beta = 0.58, 95% CI [0.17, 0.98])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 23.38 | 2.37 – 230.60 | **6.962e-03** |
| ai usage | 0.08 | 0.01 – 0.90 | **4.067e-02** |
| ai correctness | 0.21 | 0.02 – 2.16 | 1.907e-01 |
| ai usage × ai correctness | 36.67 | 2.87 – 467.98 | **5.561e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 1.22 | | |
| ICC | 0.27 | | |
| N question\_no | 70 | | |
| Observations | 280 | | |
| Marginal R2 / Conditional R2 | 0.091 / 0.338 | | |
| AIC | 241.461 | | |

EXPERT ANOVA

Data: expert\_df

Models:

expert\_model\_v1: human\_correctness ~ ai\_usage + (1 | question\_no)

expert\_model\_v2: human\_correctness ~ ai\_usage + ai\_correctness + (1 | question\_no)

expert\_model\_v3: human\_correctness ~ ai\_usage \* ai\_correctness + (1 | question\_no)

npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)

expert\_model\_v1 3 248.70 259.60 -121.35 242.70

expert\_model\_v2 4 250.08 264.62 -121.04 242.08 0.6211 1 0.430646

expert\_model\_v3 5 241.46 259.63 -115.73 231.46 10.6148 1 0.001122 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

NOVICE MODEL 1

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + (1 | question\_no)

Data: novice\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

462.3 474.4 -228.1 456.3 417

Scaled residuals:

Min 1Q Median 3Q Max

-2.7649 -0.6361 0.3617 0.5790 1.2722

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 1.014 1.007

Number of obs: 420, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.6615 0.2027 3.264 0.0011 \*\*

ai\_usage 1.1472 0.2569 4.466 7.96e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr)

ai\_usage -0.440

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage

(formula: human\_correctness ~ ai\_usage). The model included question\_no as random effect (formula: ~1 | question\_no). The

model's total explanatory power is substantial (conditional R2 = 0.29) and the part related to the fixed effects alone

(marginal R2) is of 0.07. The model's intercept, corresponding to ai\_usage = 0, is at 0.66 (95% CI [0.26, 1.06], p = 0.001).

Within this model:

- The effect of ai usage is statistically significant and positive (beta = 1.15, 95% CI [0.64, 1.65], p < .001; Std. beta =

0.57, 95% CI [0.32, 0.83])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.94 | 1.30 – 2.88 | **1.099e-03** |
| ai usage | 3.15 | 1.90 – 5.21 | **7.963e-06** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 1.01 | | |
| ICC | 0.24 | | |
| N question\_no | 70 | | |
| Observations | 420 | | |
| Marginal R2 / Conditional R2 | 0.071 / 0.290 | | |
| AIC | 462.283 | | |

NOVICE MODEL 2

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + ai\_correctness + (1 | question\_no)

Data: novice\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

459.2 475.3 -225.6 451.2 416

Scaled residuals:

Min 1Q Median 3Q Max

-2.8386 -0.6887 0.3523 0.5859 1.5477

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.8694 0.9324

Number of obs: 420, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.3622 0.4802 -0.754 0.451

ai\_usage 1.1480 0.2570 4.467 7.93e-06 \*\*\*

ai\_correctness 1.1489 0.5015 2.291 0.022 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg

ai\_usage -0.253

ai\_crrctnss -0.912 0.072

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

ai\_correctness (formula: human\_correctness ~ ai\_usage + ai\_correctness). The model included question\_no as random effect

(formula: ~1 | question\_no). The model's total explanatory power is substantial (conditional R2 = 0.29) and the part related

to the fixed effects alone (marginal R2) is of 0.10. The model's intercept, corresponding to ai\_usage = 0 and ai\_correctness

= 0, is at -0.36 (95% CI [-1.30, 0.58], p = 0.451). Within this model:

- The effect of ai usage is statistically significant and positive (beta = 1.15, 95% CI [0.64, 1.65], p < .001; Std. beta =

0.57, 95% CI [0.32, 0.83])

- The effect of ai correctness is statistically significant and positive (beta = 1.15, 95% CI [0.17, 2.13], p = 0.022; Std.

beta = 0.37, 95% CI [0.05, 0.68])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 0.70 | 0.27 – 1.78 | 4.507e-01 |
| ai usage | 3.15 | 1.90 – 5.22 | **7.929e-06** |
| ai correctness | 3.15 | 1.18 – 8.43 | **2.198e-02** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.87 | | |
| ICC | 0.21 | | |
| N question\_no | 70 | | |
| Observations | 420 | | |
| Marginal R2 / Conditional R2 | 0.100 / 0.288 | | |
| AIC | 459.150 | | |

NOVICE MODEL 3

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage \* ai\_correctness + (1 | question\_no)

Data: novice\_df

Control: glmerControl(optimizer = "bobyqa")

AIC BIC logLik deviance df.resid

459.6 479.8 -224.8 449.6 415

Scaled residuals:

Min 1Q Median 3Q Max

-2.9701 -0.6651 0.3367 0.5880 1.2820

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.8802 0.9382

Number of obs: 420, groups: question\_no, 70

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.0005881 0.5579703 0.001 0.999

ai\_usage 0.4038007 0.6376223 0.633 0.527

ai\_correctness 0.7370771 0.5962007 1.236 0.216

ai\_usage:ai\_correctness 0.8801030 0.6955654 1.265 0.206

Correlation of Fixed Effects:

(Intr) ai\_usg a\_crrc

ai\_usage -0.563

ai\_crrctnss -0.936 0.529

a\_sg:\_crrct 0.516 -0.914 -0.547

We fitted a logistic mixed model (estimated using ML and BOBYQA optimizer) to predict human\_correctness with ai\_usage and

ai\_correctness (formula: human\_correctness ~ ai\_usage \* ai\_correctness). The model included question\_no as random effect

(formula: ~1 | question\_no). The model's total explanatory power is substantial (conditional R2 = 0.30) and the part related

to the fixed effects alone (marginal R2) is of 0.11. The model's intercept, corresponding to ai\_usage = 0 and ai\_correctness

= 0, is at 5.88e-04 (95% CI [-1.09, 1.09], p > .999). Within this model:

- The effect of ai usage is statistically non-significant and positive (beta = 0.40, 95% CI [-0.85, 1.65], p = 0.527; Std.

beta = 0.59, 95% CI [0.34, 0.85])

- The effect of ai correctness is statistically non-significant and positive (beta = 0.74, 95% CI [-0.43, 1.91], p = 0.216;

Std. beta = 0.37, 95% CI [0.06, 0.69])

- The effect of ai usage × ai correctness is statistically non-significant and positive (beta = 0.88, 95% CI [-0.48, 2.24],

p = 0.206; Std. beta = 0.14, 95% CI [-0.08, 0.36])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.00 | 0.34 – 2.99 | 9.992e-01 |
| ai usage | 1.50 | 0.43 – 5.23 | 5.265e-01 |
| ai correctness | 2.09 | 0.65 – 6.72 | 2.164e-01 |
| ai usage × ai correctness | 2.41 | 0.62 – 9.43 | 2.058e-01 |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.88 | | |
| ICC | 0.21 | | |
| N question\_no | 70 | | |
| Observations | 420 | | |
| Marginal R2 / Conditional R2 | 0.109 / 0.297 | | |
| AIC | 459.564 | | |

NOVICE ANOVA

Data: novice\_df

Models:

novice\_model\_v1: human\_correctness ~ ai\_usage + (1 | question\_no)

novice\_model\_v2: human\_correctness ~ ai\_usage + ai\_correctness + (1 | question\_no)

novice\_model\_v3: human\_correctness ~ ai\_usage \* ai\_correctness + (1 | question\_no)

npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)

novice\_model\_v1 3 462.28 474.40 -228.14 456.28

novice\_model\_v2 4 459.15 475.31 -225.57 451.15 5.1324 1 0.02348 \*

novice\_model\_v3 5 459.56 479.76 -224.78 449.56 1.5866 1 0.20781

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

CORRECT MODEL V1

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + seniority + (1 | question\_no)

Data: correct\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

594.0 611.8 -293.0 586.0 616

Scaled residuals:

Min 1Q Median 3Q Max

-3.6333 0.1817 0.3308 0.4661 1.3311

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.7917 0.8898

Number of obs: 620, groups: question\_no, 62

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.4354 0.2390 6.005 1.91e-09 \*\*\*

ai\_usage 1.1985 0.2285 5.246 1.56e-07 \*\*\*

seniorityNovice -0.6855 0.2308 -2.970 0.00298 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg

ai\_usage -0.257

seniortyNvc -0.651 -0.068

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

seniority (formula: human\_correctness ~ ai\_usage + seniority). The model included question\_no as random effect (formula: ~1

| question\_no). The model's total explanatory power is substantial (conditional R2 = 0.28) and the part related to the fixed

effects alone (marginal R2) is of 0.10. The model's intercept, corresponding to ai\_usage = 0 and seniority = Expert, is at

1.44 (95% CI [0.97, 1.90], p < .001). Within this model:

- The effect of ai usage is statistically significant and positive (beta = 1.20, 95% CI [0.75, 1.65], p < .001; Std. beta =

0.60, 95% CI [0.38, 0.82])

- The effect of seniority [Novice] is statistically significant and negative (beta = -0.69, 95% CI [-1.14, -0.23], p =

0.003; Std. beta = -0.69, 95% CI [-1.14, -0.23])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 4.20 | 2.63 – 6.71 | **1.915e-09** |
| ai usage | 3.32 | 2.12 – 5.19 | **1.558e-07** |
| seniority [Novice] | 0.50 | 0.32 – 0.79 | **2.980e-03** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.79 | | |
| ICC | 0.19 | | |
| N question\_no | 62 | | |
| Observations | 620 | | |
| Marginal R2 / Conditional R2 | 0.104 / 0.278 | | |
| AIC | 594.040 | | |

CORRECT\_MODEL\_V2

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + (1 | question\_no)

Data: correct\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

601.3 614.6 -297.7 595.3 617

Scaled residuals:

Min 1Q Median 3Q Max

-3.4778 0.2329 0.3465 0.5166 1.3166

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.7432 0.8621

Number of obs: 620, groups: question\_no, 62

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.9974 0.1763 5.658 1.53e-08 \*\*\*

ai\_usage 1.1720 0.2222 5.275 1.33e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr)

ai\_usage -0.406

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage

(formula: human\_correctness ~ ai\_usage). The model included question\_no as random effect (formula: ~1 | question\_no). The

model's total explanatory power is moderate (conditional R2 = 0.25) and the part related to the fixed effects alone

(marginal R2) is of 0.08. The model's intercept, corresponding to ai\_usage = 0, is at 1.00 (95% CI [0.65, 1.34], p < .001).

Within this model:

- The effect of ai usage is statistically significant and positive (beta = 1.17, 95% CI [0.74, 1.61], p < .001; Std. beta =

0.59, 95% CI [0.37, 0.80])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 2.71 | 1.92 – 3.83 | **1.533e-08** |
| ai usage | 3.23 | 2.09 – 4.99 | **1.330e-07** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.74 | | |
| ICC | 0.18 | | |
| N question\_no | 62 | | |
| Observations | 620 | | |
| Marginal R2 / Conditional R2 | 0.079 / 0.248 | | |
| AIC | 601.320 | | |

CORRECT ANOVA

Data: correct\_df

Models:

correct\_model\_v2: human\_correctness ~ ai\_usage + (1 | question\_no)

correct\_model\_v1: human\_correctness ~ ai\_usage + seniority + (1 | question\_no)

npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)

correct\_model\_v2 3 601.32 614.61 -297.66 595.32

correct\_model\_v1 4 594.04 611.76 -293.02 586.04 9.2805 1 0.002316 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

WRONG MODEL V1

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + seniority + (1 | question\_no)

Data: wrong\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

103.4 112.9 -47.7 95.4 76

Scaled residuals:

Min 1Q Median 3Q Max

-2.1463 -0.8070 0.4168 0.6519 1.6792

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.8219 0.9066

Number of obs: 80, groups: question\_no, 8

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.6978 0.6466 2.626 0.00864 \*\*

ai\_usage -0.4031 0.5191 -0.777 0.43744

seniorityNovice -1.2893 0.5655 -2.280 0.02261 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) ai\_usg

ai\_usage -0.456

seniortyNvc -0.635 0.054

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage and

seniority (formula: human\_correctness ~ ai\_usage + seniority). The model included question\_no as random effect (formula: ~1

| question\_no). The model's total explanatory power is substantial (conditional R2 = 0.28) and the part related to the fixed

effects alone (marginal R2) is of 0.10. The model's intercept, corresponding to ai\_usage = 0 and seniority = Expert, is at

1.70 (95% CI [0.43, 2.97], p = 0.009). Within this model:

- The effect of ai usage is statistically non-significant and negative (beta = -0.40, 95% CI [-1.42, 0.61], p = 0.437; Std.

beta = -0.20, 95% CI [-0.71, 0.31])

- The effect of seniority [Novice] is statistically significant and negative (beta = -1.29, 95% CI [-2.40, -0.18], p =

0.023; Std. beta = -1.29, 95% CI [-2.40, -0.18])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 5.46 | 1.54 – 19.40 | **8.644e-03** |
| ai usage | 0.67 | 0.24 – 1.85 | 4.374e-01 |
| seniority [Novice] | 0.28 | 0.09 – 0.83 | **2.261e-02** |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.82 | | |
| ICC | 0.20 | | |
| N question\_no | 8 | | |
| Observations | 80 | | |
| Marginal R2 / Conditional R2 | 0.098 / 0.278 | | |
| AIC | 103.419 | | |

WRONG MODEL V2

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']

Family: binomial ( logit )

Formula: human\_correctness ~ ai\_usage + (1 | question\_no)

Data: wrong\_df

Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "L-BFGS-B"))

AIC BIC logLik deviance df.resid

107.1 114.3 -50.6 101.1 77

Scaled residuals:

Min 1Q Median 3Q Max

-1.8546 -0.9692 0.5392 0.7077 1.2414

Random effects:

Groups Name Variance Std.Dev.

question\_no (Intercept) 0.6501 0.8063

Number of obs: 80, groups: question\_no, 8

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.8406 0.4658 1.805 0.0711 .

ai\_usage -0.3699 0.4967 -0.745 0.4564

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr)

ai\_usage -0.564

We fitted a logistic mixed model (estimated using ML and optimx optimizer) to predict human\_correctness with ai\_usage

(formula: human\_correctness ~ ai\_usage). The model included question\_no as random effect (formula: ~1 | question\_no). The

model's total explanatory power is moderate (conditional R2 = 0.17) and the part related to the fixed effects alone

(marginal R2) is of 8.72e-03. The model's intercept, corresponding to ai\_usage = 0, is at 0.84 (95% CI [-0.07, 1.75], p =

0.071). Within this model:

- The effect of ai usage is statistically non-significant and negative (beta = -0.37, 95% CI [-1.34, 0.60], p = 0.456; Std.

beta = -0.19, 95% CI [-0.68, 0.30])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence

Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **human\_correctness** | | |
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 2.32 | 0.93 – 5.78 | 7.113e-02 |
| ai usage | 0.69 | 0.26 – 1.83 | 4.564e-01 |
| **Random Effects** | | | |
| σ2 | 3.29 | | |
| τ00 question\_no | 0.65 | | |
| ICC | 0.17 | | |
| N question\_no | 8 | | |
| Observations | 80 | | |
| Marginal R2 / Conditional R2 | 0.009 / 0.172 | | |
| AIC | 107.130 | | |

WRONG ANOVA

Data: wrong\_df

Models:

wrong\_model\_v2: human\_correctness ~ ai\_usage + (1 | question\_no)

wrong\_model\_v1: human\_correctness ~ ai\_usage + seniority + (1 | question\_no)

npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)

wrong\_model\_v2 3 107.13 114.28 -50.565 101.130

wrong\_model\_v1 4 103.42 112.95 -47.709 95.419 5.7108 1 0.01686 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1