Study Detailed Results

Contents

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

S11.1 Statistcs

These statistics are derived from programmatically generated study instance-specific stats packs located at doi:10.21954/ou.rd.28044944 [path: /study/data/stats/generated]. These have been combined into consolidated cross-instance Microsoft Excel spreadsheets, from which the tables and figures here have been taken. These are located at: doi:10.21954/ou.rd.28044944 [path: /study/data/stats/combined].

S11.1.1 Classification Agreement

Table S11.1 shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant for each study instance. Rows are broken down by UD-ML model, with an ALL row containing data across all models. Note that not all UD-ML models were included in each study – each row shows aggregated data for a model for all studies that contained that model. The second column indicates the number of studies that each model appeared in.

Percent of Classification Manual Agrees

This table shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This data only includes items for which the participant entered classification values during the training phase of the study. It is consolidated across all personas - some models feature in more than one persona study, as indicated in the '#' column. The ALL column is caluclated as a mean of the individual persona values.

		Susan			Adam			Phoebe			Kenton			Usha			ALL				
Model	**#**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**		
ALL		4,025	3,648	90.6%	12,021	11,344	94.4%	3876	3535	91.2%	964	908	94.2%	5,157	4,726	91.6%	26,043	24,161	92.8%		ALL
urgency	5	503	499	99.2%	1,277	1,259	98.6%	646	609	94.3%	135	133	98.5%	573	540	94.2%	3,134	3,040	97.0%	*+4.2*	ur
work-logistics	5	503	482	95.8%	1,277	1,250	97.9%	646	622	96.3%	135	131	97.0%	573	545	95.1%	3,134	3,030	96.7%	*+3.9*	w-l
work-pers	5	503	423	84.1%	1,277	1,210	94.8%	646	586	90.7%	135	121	89.6%	573	511	89.2%	3,134	2,851	91.0%	*-1.8*	w-p
interested	4	504	361	71.6%	1,277	1,186	92.9%				136	122	89.7%	573	462	80.6%	2,490	2,131	85.6%	*-7.2*	in
work-relevant	4	503	462	91.8%	1,277	1,134	88.8%	646	576	89.2%				573	527	92.0%	2,999	2,699	90.0%	*-2.8*	w-r
company-law	1													573	539	94.1%	573	539	94.1%	*+1.3*	co-l
cycling	1				1,277	1,180	92.4%										1,277	1,180	92.4%	*-0.4*	cy
cycling-logistics	1				529	492	93.0%										529	492	93.0%	*+0.2*	cy-l
football	1										141	127	90.1%				141	127	90.1%	*-2.7*	fo
friend-group	1							646	575	89.0%							646	575	89.0%	*-3.8*	f-g
golf	1										141	139	98.6%				141	139	98.6%	*+5.8*	go
golf-logistics	1										141	135	95.7%				141	135	95.7%	*+3.0*	g-l
pers-urgency	1				1,276	1,208	94.7%										1,276	1,208	94.7%	*+1.9*	p-u
personal-interested	1							646	567	87.8%							646	567	87.8%	*-5.0*	p-i
riding	1													573	526	91.8%	573	526	91.8%	*-1.0*	ri
riding-arrangements	1													573	558	97.4%	573	558	97.4%	*+4.6*	r-a
school-importance	1													573	518	90.4%	573	518	90.4%	*-2.4*	s-i
tech	1				1,277	1,225	95.9%										1,277	1,225	95.9%	*+3.2*	tc
tennis	1	503	445	88.5%													503	445	88.5%	*-4.3*	tn
tennis-arrangements	1	503	490	97.4%													503	490	97.4%	*+4.6*	t-a
tennis-organising	1	503	486	96.6%													503	486	96.6%	*+3.8*	t-o
work-urgency	1				1,277	1,200	94.0%										1,277	1,200	94.0%	*+1.2*	w-u

count_items Number of items in the dataset 'phase2' having a manual classification record (the numerator in the percentage calculation)
count_agree Number of items in the dataset 'phase2' having a manual classification record that agrees with the ML classification (the denominator in the percentage calculation)
percentage_agree Percentage of records in dataset 'phase2' having a manual classification, which agrees with the ML classification

Table S11.1: Percent of Manual Classification Agreement with UD-ML

Table S11.2 and Figure S11.1 show how the classification agreement percentage changed over time during the training phase of each study, which occurred over a period of up to 7 days for each study.

Percent of Classification Manual Agrees (Time Series)

This table shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This consolidated time series data shows how the measure changed over time during the study. This data only includes items for which the participant-entered classification values during the training phase of the study.

	Susan			Adam			Phoebe			Kenton			Usha			ALL		
Day	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**	**Items**	**Agree**	**% Agree**
Day 1	472	417	88.3%	1,359	1,244	91.5%	366	338	92.3%	204	189	92.6%	342	307	89.8%	2,743	2,495	91.0%
Day 2	600	539	89.8%	1,547	1,428	92.3%	420	363	86.4%	266	253	95.1%	522	459	87.9%	3,355	3,042	90.7%
Day 3	776	692	89.2%	2,205	2,089	94.7%	426	387	90.8%	98	90	91.8%	747	662	88.6%	4,252	3,920	92.2%
Day 4	72	68	94.4%	1,830	1,736	94.9%	132	118	89.4%	98	91	92.9%	1,080	995	92.1%	3,212	3,008	93.6%
Day 5	672	608	90.5%	2,680	2,532	94.5%	852	773	90.7%	258	247	95.7%	675	615	91.1%	5,137	4,775	93.0%
Day 6	1,024	947	92.5%	920	890	96.7%	1,392	1,285	92.3%	37	36	97.3%	693	651	93.9%	4,066	3,809	93.7%
Day 7	409	377	92.2%	1,440	1,385	96.2%	288	271	94.1%	3	2	66.7%	900	851	94.6%	3,040	2,886	94.9%

Table S11.2: Percent of Manual Classification Agreement with UD-ML time Series

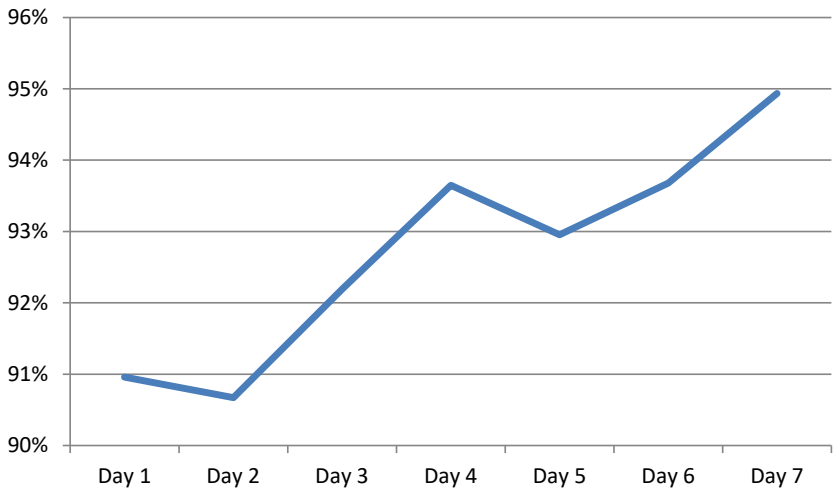


Figure S11.1: Percent of Manual Classification Agreement with UD-ML Time Series

Table S11.3 and Figure S11.2 show the classification agreement data as a Cohen's Kappa.

Cohen's Kappa For Manual Classification Agreement

This table shows Cohen's Kappa for agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This data only includes items for which the participant-entered classification values during the training phase of the study. It is consolidated across all personas - some models feature in more than one persona study, as indicated in the '#' column. The ALL column is caluclated as a mean of the individual persona values. The ALL row was calculated separately for each persona; it is not a mean of the individual model values.

Model	#	Susan	Adam	Phoebe	Kenton	Usha	ALL		
ALL		0.756	0.881	0.827	0.871	0.818	0.83		ALL
urgency	5	0.663	0.797	0.596	0.881	0.607	0.71	*-0.12*	ur
work-logistics	5	0.379	0.856	0.687	0.784	0.475	0.64	*-0.19*	w-l
work-pers	5	0.639	0.865	0.792	0.708	0.736	0.75	*-0.08*	w-p
interested	4	0.561	0.838		0.772	0.687	0.71	*-0.12*	in
work-relevant	4	0.000	0.462	0.68		0.766	0.48	*-0.35*	w-r
company-law	1					0.606	0.61	*-0.22*	co-l
cycling	1		0.790				0.79	*-0.04*	cy
cycling-logistics	1		0.555				0.56	*-0.28*	cy-l
football	1				0.791		0.79	*-0.04*	fo
friend-group	1			0.743			0.74	*-0.09*	f-g
golf	1				0.867		0.87	*+0.04*	go
golf-logistics	1				0.549		0.55	*-0.28*	g-l
pers-urgency	1		0.748				0.75	*-0.08*	p-u
personal-interested	1			0.696			0.70	*-0.13*	p-i
riding	1					0.725	0.73	*-0.11*	ri
riding-arrangements	1					0.638	0.64	*-0.19*	r-a
school-importance	1					0.427	0.43	*-0.40*	s-i
tech	1		0.765				0.77	*-0.07*	tc
tennis	1	0.475					0.48	*-0.36*	tn
tennis-arrangements	1	0.506					0.51	*-0.32*	t-a
tennis-organising	1	0.306					0.31	*-0.52*	t-o
work-urgency	1		0.642				0.64	*-0.19*	w-u

Note that the ALL column represents a MEAN of the Cohen's Kappa values for each of the Persona columns

Table S11.3: Cohen's Kappa for Manual Classification Agreement with UD-ML

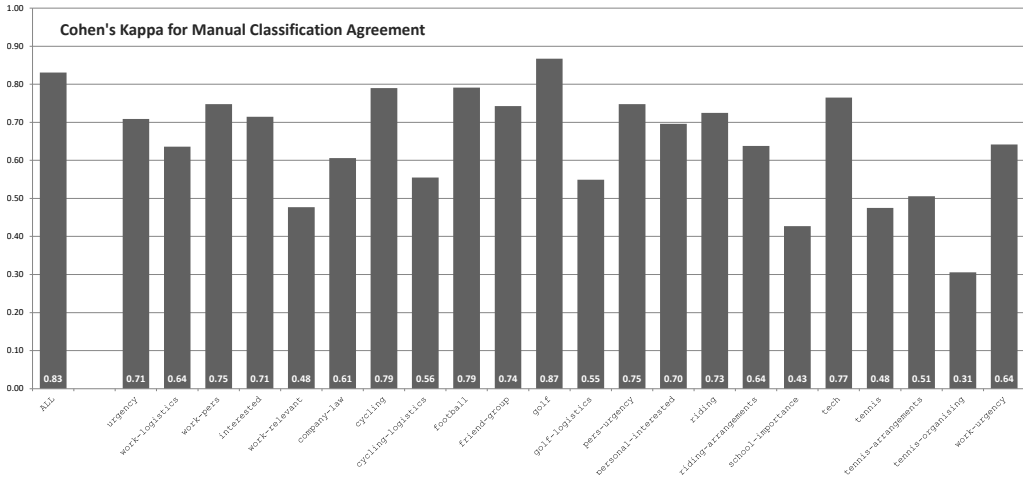


Figure S11.2: Cohen's Kappa for Manual Classification Agreement with UD-ML

S11.1.2 Evaluation Feedback

Table S11.4 details the Pearson's correlation coefficient (r) between synthetic evaluations and participant-entered feedback, broken down by Persona, UD-ML Model and Evaluation Tag for Mode 1 and Mode 2 evaluations. Mean values of r have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to r. Only combinations of data having a significant correlation ($p \leq 0.06$) value are included in the mean.

Pearson Correlation for Evaluation Feedback

This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the correlation between synthetic evaluations and participant-entered feedback. The higher the correlation, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those r values having a significant correlation (p < 0.06); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 1 & 2 evaluations have feedback.

Persona	Model	vanilla-mode1-01		vanilla-mode1-02		vanilla-mode2-01		base-mode2-01		vanilla4-mode2-01		vanilla4o-mode2-01		Mean r		
		r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*			
Susan	ALL	0.11	0.051	0.08	0.047	0.02	0.806	0.28	0.000	0.54	0.000	0.73	0.000	**0.38**		ALL
Susan	interested	0.44	0.004	0.07	0.545	0.25	0.161	0.38	0.044	0.54	0.005	0.63	0.000	**0.50**	+0.13	in
Susan	tennis	0.38	0.011	0.16	0.149	-0.26	0.156	0.12	0.519	0.33	0.113	0.73	0.000	**0.58**	+0.21	tn
Susan	tennis-arrangements	0.20	0.374	-0.20	0.058	0.27	0.140	0.10	0.596	0.52	0.007	0.49	0.000	**0.29**	-0.08	t-a
Susan	tennis-organising	-0.21	0.170	-0.13	0.226	-0.13	0.483	0.24	0.228	0.48	0.012	0.66	0.000	**0.58**	+0.20	t-o
Susan	urgency	0.05	0.730	0.24	0.022	-0.07	0.711	0.00	0.000	0.41	0.037	0.85	0.000	**0.57**	+0.19	ur
Susan	work-logistics	0.14	0.371	0.30	0.005	0.02	0.933	-0.07	0.734	0.40	0.043	0.97	0.000	**0.74**	+0.36	w-l
Susan	work-pers	0.16	0.312	-0.01	0.956	0.19	0.311	0.10	0.605	0.60	0.001	0.74	0.000	**0.68**	+0.30	w-p
Susan	work-relevant	-0.17	0.266	0.24	0.025	-0.33	0.065	0.56	0.002	0.78	0.000	0.70	0.000	**0.45**	+0.08	w-r
Adam	ALL	0.36	0.000	0.20	0.001	0.08	0.137	0.35	0.000	0.65	0.000	0.66	0.000	**0.46**		ALL
Adam	cycling	0.16	0.395	0.61	0.001	0.12	0.487	0.18	0.302	0.64	0.000	0.44	0.043	**0.57**	+0.10	cy
Adam	cycling-logistics	0.40	0.145	0.56	0.020	0.13	0.583	-0.13	0.568	0.49	0.011	0.60	0.009	**0.55**	+0.08	c-l
Adam	interested	0.57	0.001	0.54	0.002	0.36	0.031	0.31	0.068	0.61	0.000	0.84	0.000	**0.57**	+0.10	in
Adam	pers-urgency	0.35	0.058	0.13	0.497	0.07	0.666	0.31	0.070	0.57	0.000	0.61	0.003	**0.47**	+0.01	p-u
Adam	tech	-0.11	0.560	0.19	0.321	-0.05	0.750	0.71	0.000	0.96	0.000	0.69	0.000	**0.84**	+0.38	tc
Adam	urgency	0.42	0.020	0.00	0.988	-0.07	0.688	0.53	0.001	0.35	0.026	0.49	0.022	**0.45**	-0.02	ur
Adam	work-logistics	0.81	0.000	0.25	0.204	-0.10	0.583	0.56	0.000	0.71	0.000	0.51	0.014	**0.66**	+0.20	w-l
Adam	work-pers	-0.07	0.722	0.38	0.048	-0.12	0.494	0.43	0.015	0.68	0.000	0.64	0.002	**0.55**	+0.08	w-p
Adam	work-relevant	0.40	0.028	-0.09	0.636	0.07	0.695	0.52	0.001	0.71	0.000	0.92	0.000	**0.70**	+0.23	w-r
Adam	work-urgency	0.13	0.491	-0.18	0.341	-0.07	0.696	0.25	0.136	0.85	0.000	0.94	0.000	**0.90**	+0.44	w-u
Phoebe	ALL									0.79	0.000	0.88	0.000	**0.84**		ALL
Phoebe	friend-group									0.66	0.002	0.81	0.000	**0.75**	-0.09	f-g
Phoebe	personal-interested									0.88	0.000	0.61	0.001	**0.78**	-0.06	p-i
Phoebe	urgency									0.53	0.016	0.69	0.000	**0.62**	-0.22	ur
Phoebe	work-logistics									0.80	0.000	0.90	0.000	**0.86**	+0.02	w-l
Phoebe	work-pers									0.85	0.000	0.90	0.000	**0.88**	+0.04	w-p
Phoebe	work-relevant									0.83	0.000	0.96	0.000	**0.91**	+0.07	w-r
Kenton	ALL	0.24	0.000	0.20	0.000	0.06	0.278	0.34	0.000	0.84	0.000	0.73	0.000	**0.53**		ALL
Kenton	football	0.41	0.001	0.29	0.001	-0.06	0.676	0.28	0.067	1.00	0.000	0.74	0.000	**0.59**	+0.06	fo
Kenton	golf	-0.06	0.621	-0.07	0.429	0.02	0.872	0.51	0.000	0.00	0.000	0.28	0.014	**0.40**	-0.12	go
Kenton	golf-logistics	0.00	0.975	0.21	0.021	0.04	0.774	0.47	0.002	0.00	0.000	0.21	0.074	**0.34**	-0.19	g-l
Kenton	interested	0.30	0.016	0.25	0.005	0.10	0.523	0.14	0.379	0.78	0.000	0.87	0.000	**0.62**	+0.09	in
Kenton	urgency	0.50	0.000	0.37	0.000	-0.11	0.468	0.60	0.000	0.72	0.000	0.69	0.000	**0.59**	+0.06	ur
Kenton	work-logistics	0.38	0.003	0.11	0.233	-0.02	0.908	0.01	0.963	0.67	0.000	0.86	0.000	**0.68**	+0.15	w-l
Kenton	work-pers	0.14	0.267	0.26	0.003	0.01	0.970	0.19	0.288	0.94	0.000	0.72	0.000	**0.75**	+0.22	w-p
Usha	ALL									0.70	0.000	0.82	0.000	**0.77**		ALL
Usha	company-law									0.40	0.008	0.87	0.000	**0.71**	-0.06	c-l
Usha	interested									0.79	0.000	0.74	0.000	**0.76**	-0.01	in
Usha	riding									0.91	0.000	0.79	0.000	**0.86**	+0.09	ri
Usha	riding-arrangements									0.83	0.000	0.83	0.000	**0.83**	+0.06	r-a
Usha	school-importance									0.30	0.054	0.81	0.000	**0.61**	-0.16	s-i
Usha	urgency									0.79	0.000	0.75	0.000	**0.77**	-0.00	ur
Usha	work-logistics									0.90	0.000	0.79	0.000	**0.85**	+0.08	w-l
Usha	work-pers									0.60	0.000	0.92	0.000	**0.81**	+0.04	w-p
Usha	work-relevant									0.83	0.000	0.75	0.000	**0.79**	+0.02	w-r

Model label:	V-M1-A	V-M1-B	V-M2	B-M2	V4-M2	V4o-M2
Percentage of significant rows:	31%	36%	4%	33%	93%	98%
Mean r (significant only, excl. ALL):	**0.45**	**0.31**	**0.01**	**0.49**	**0.68**	**0.74**
Mean r (significant only, ALL only):	**0.24**	**0.16**	**n/a**	**0.32**	**0.70**	**0.76**

Table S11.4: r for Evaluation Feedback by Persona, UD-ML Model and Evaluation Tag

Figure S11.3 shows the mean correlation (r) between Modes 1 and 2 synthetic evaluation and participant feedback for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.4.

Figure S11.4 shows the percentages of rows in Table S11.4 for each evaluation tag having a significant correlation ($p \leq 0.06$). Only these row values are used to calculate the mean data that is illustrated in Figures S11.5.

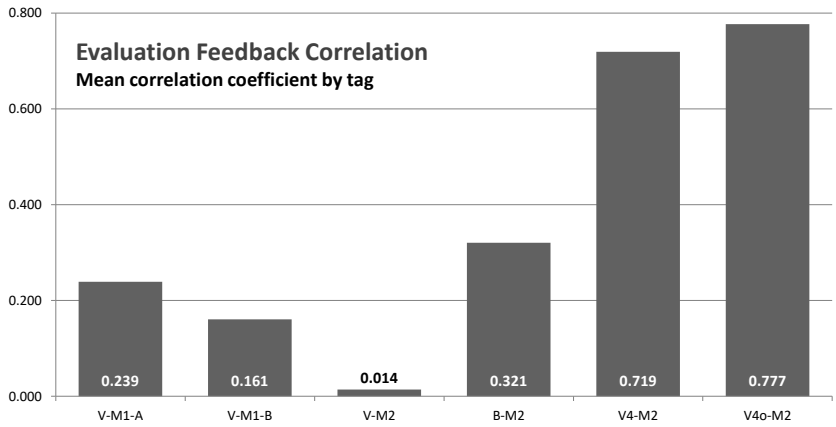


Figure S11.3: Mean Synthetic *vs* Participant Evaluation r by Evaluation Tag

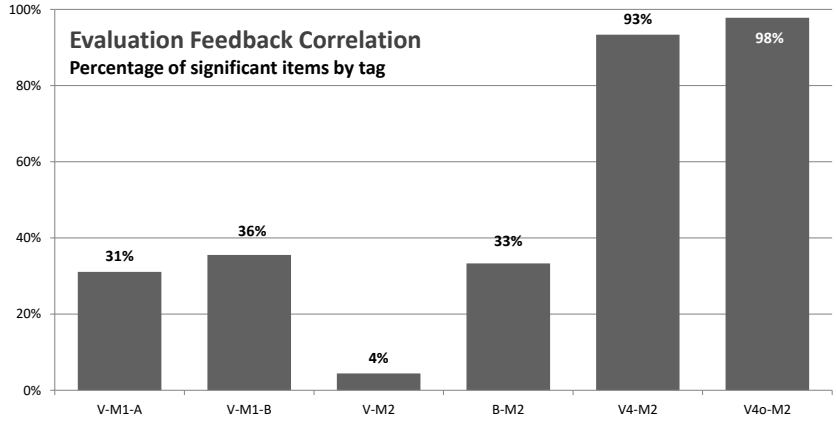
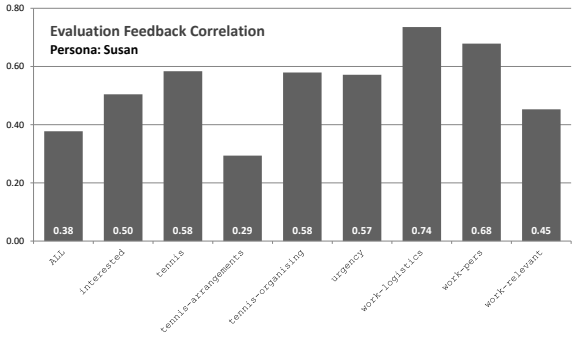
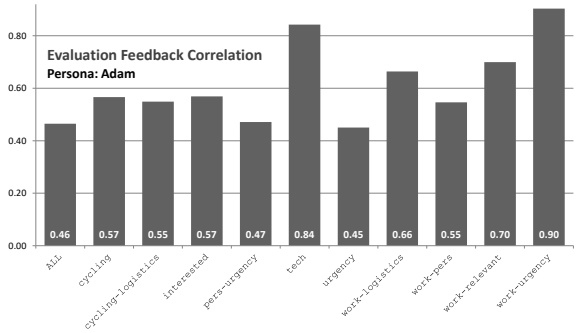


Figure S11.4: Percentage of Items Having Significant Correlation (r) by Evaluation Tag

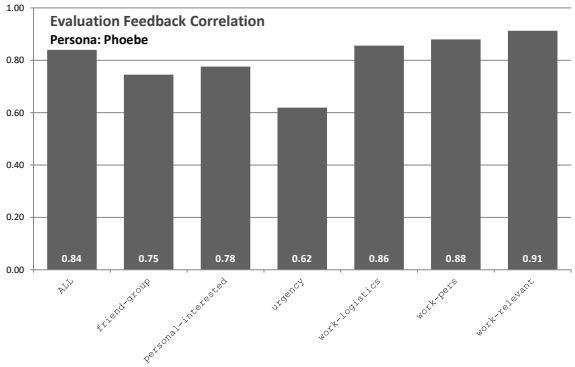
Figures S11.5a to S11.5e show the mean correlation (r) between Modes 1 and 2 synthetic evaluation and participant feedback by UD-ML Model for each persona; this data is sourced from Table S11.4.



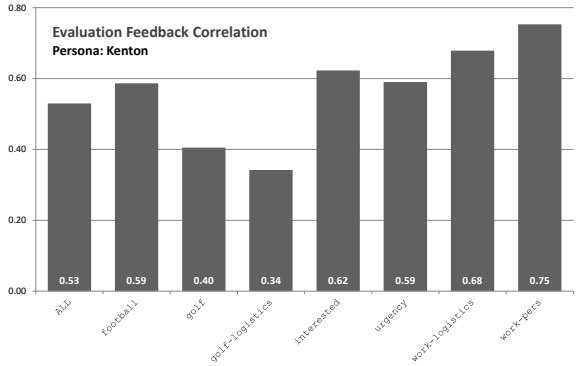
(a) Persona: Susan



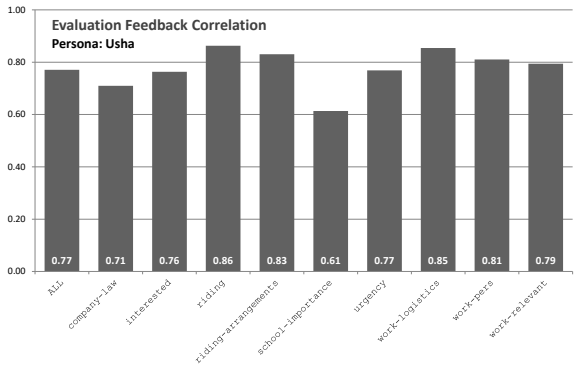
(b) Persona: Adam



(c) Persona: Phoebe



(d) Persona: Kenton



(e) Persona: Usha

Figure S11.5: Synthetic *vs* Participant Evaluation r by UD-ML Model

S11.1.3 Evaluation Agreement (Modes 1 & 2)

Table S11.5 details the point-biserial correlation coefficient (r_{pb}) between Mode 1 and Mode 2 synthetic evaluations and classification actions performed by the study participant during the training phase. In this case, r_{pb} is used because Mode 1 and Mode 2 evaluations have a categorical value (1-5 on the Likert scale) while classification actions are dichotomous (agree/disagree). The data is broken down by Persona, UD-ML Model and Evaluation Tag. Mean values of r_{pb} have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to r_{pb}. Only combinations of data having a significant correlation ($p \leq 0.06$) value are included in the mean.

Point Biserial Correlation for Evaluation Agreement

This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the Point Biserial correlation between synthetic evaluations and classification actions entered by the participant. The higher the correlation, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those r values having a significant correlation (p < 0.06); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 1 & 2 evaluations have feedback, and this data only includes items for which the participant entered classification values during the training phase of the study.

Persona	Model	vanilla-mode1-01		vanilla-mode1-02		vanilla-mode2-01		base-mode2-01		vanilla4-mode2-01		vanilla4o-mode2-01		Mean r		
		rbp	*p*	*rbp*	*p*	*rbp*	*p*	*rbp*	*p*	*rbp*	*p*	*rbp*	*p*			
Susan	ALL	0.12	0.000	0.03	0.282	0.00	0.989	0.11	0.002	0.36	0.000	0.50	0.000	**0.28**		ALL
Susan	interested	0.12	0.128	0.02	0.796	0.03	0.718	-0.01	0.937	0.32	0.001	0.38	0.000	**0.35**	+0.07	in
Susan	tennis	0.10	0.192	0.15	0.064	-0.15	0.048	0.07	0.472	0.46	0.000	0.48	0.000	**0.25**	-0.03	tn
Susan	tennis-arrangements	0.09	0.321	-0.10	0.179	-0.01	0.878	0.21	0.041	0.38	0.000	0.38	0.000	**0.32**	+0.04	t-a
Susan	tennis-organising	0.02	0.809	-0.02	0.798	-0.03	0.725	-0.03	0.774	-0.05	0.609	0.72	0.000	**0.72**	+0.44	t-o
Susan	urgency	-0.06	0.449	-0.04	0.593	-0.07	0.379	0.00	0.000	0.00	0.000	0.17	0.045	**0.17**	-0.11	ur
Susan	work-logistics	0.21	0.007	0.21	0.011	-0.03	0.729	-0.04	0.682	0.14	0.170	0.63	0.000	**0.37**	+0.09	w-l
Susan	work-pers	0.14	0.075	0.00	0.964	0.12	0.112	-0.23	0.030	0.27	0.008	0.38	0.000	**0.15**	-0.13	w-p
Susan	work-relevant	-0.02	0.830	0.14	0.080	-0.10	0.189	0.38	0.000	0.38	0.000	0.54	0.000	**0.43**	+0.15	w-r
Adam	ALL	0.14	0.000	0.04	0.018	0.00	0.949	0.12	0.000	0.24	0.000	0.41	0.000	**0.19**		ALL
Adam	cycling	-0.05	0.422	0.06	0.220	0.11	0.088	0.35	0.000	0.47	0.000	0.47	0.000	**0.43**	+0.24	cy
Adam	cycling-logistics	0.27	0.008	0.24	0.001	0.14	0.169	-0.14	0.350	0.32	0.001	0.45	0.000	**0.32**	+0.13	c-l
Adam	interested	0.27	0.000	0.14	0.003	0.17	0.010	0.11	0.265	0.18	0.005	0.29	0.000	**0.21**	+0.02	in
Adam	pers-urgency	0.26	0.000	0.01	0.767	-0.06	0.380	0.15	0.127	0.09	0.181	0.41	0.000	**0.33**	+0.14	p-u
Adam	tech	0.05	0.468	0.01	0.788	-0.05	0.454	0.13	0.190	0.58	0.000	0.68	0.000	**0.63**	+0.44	tc
Adam	urgency	0.11	0.087	0.05	0.268	-0.01	0.935	0.32	0.001	-0.04	0.583	0.16	0.002	**0.24**	+0.05	ur
Adam	work-logistics	0.18	0.006	-0.05	0.260	-0.09	0.200	-0.02	0.818	0.37	0.000	0.49	0.000	**0.35**	+0.16	w-l
Adam	work-pers	0.03	0.682	0.05	0.298	-0.03	0.676	-0.08	0.433	0.21	0.001	0.44	0.000	**0.33**	+0.14	w-p
Adam	work-relevant	0.20	0.002	0.10	0.026	-0.01	0.916	0.21	0.031	0.16	0.014	0.36	0.000	**0.21**	+0.02	w-r
Adam	work-urgency	0.10	0.118	-0.04	0.418	-0.16	0.018	0.10	0.330	0.21	0.001	0.39	0.000	**0.15**	-0.04	w-u
Phoebe	ALL									0.33	0.000	0.45	0.000	**0.39**		ALL
Phoebe	friend-group									0.29	0.000	0.40	0.000	**0.34**	-0.05	f-g
Phoebe	personal-interested									0.25	0.001	0.32	0.000	**0.28**	-0.11	p-i
Phoebe	urgency									0.25	0.001	0.47	0.000	**0.36**	-0.03	ur
Phoebe	work-logistics									0.25	0.001	0.50	0.000	**0.38**	-0.01	w-l
Phoebe	work-pers									0.44	0.000	0.46	0.000	**0.45**	+0.06	w-p
Phoebe	work-relevant									0.44	0.000	0.56	0.000	**0.50**	+0.11	w-r
Kenton	ALL	-0.03	0.653	-0.01	0.919	-0.10	0.120	0.14	0.002	0.46	0.000	0.58	0.000	**0.41**		ALL
Kenton	football	-0.13	0.473	0.21	0.254	-0.21	0.242	0.33	0.003	0.63	0.000	0.86	0.000	**0.66**	+0.25	fo
Kenton	golf	0.27	0.129	-0.02	0.932	0.13	0.461	0.16	0.176	0.70	0.000	0.83	0.000	**0.78**	+0.37	go
Kenton	golf-logistics	0.03	0.889	0.29	0.114	0.02	0.901	0.42	0.000	1.00	0.000	0.89	0.000	**0.74**	+0.34	g-l
Kenton	interested	-0.05	0.770	-0.10	0.585	-0.11	0.544	-0.07	0.573	0.22	0.156	0.32	0.004	**0.32**	-0.09	in
Kenton	urgency	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.57	0.000	0.36	0.001	**0.47**	+0.06	ur
Kenton	work-logistics	0.51	0.003	-0.13	0.500	-0.22	0.227	-0.04	0.709	0.39	0.009	0.46	0.000	**0.46**	+0.05	w-l
Kenton	work-pers	-0.36	0.047	-0.15	0.430	-0.41	0.022	-0.04	0.726	0.44	0.003	0.53	0.000	**0.06**	-0.34	w-p
Usha	ALL									0.40	0.000	0.50	0.000	**0.45**		ALL
Usha	company-law									0.26	0.022	0.63	0.000	**0.47**	+0.01	c-l
Usha	interested									0.34	0.002	0.42	0.000	**0.38**	-0.07	in
Usha	riding									0.53	0.000	0.62	0.000	**0.57**	+0.12	ri
Usha	riding-arrangements									-0.04	0.755	-0.04	0.706	**0.00**	-0.45	r-a
Usha	school-importance									0.44	0.000	0.46	0.000	**0.45**	-0.00	s-i
Usha	urgency									0.06	0.631	0.56	0.000	**0.56**	+0.10	ur
Usha	work-logistics									0.27	0.017	0.45	0.000	**0.37**	-0.09	w-l
Usha	work-pers									0.52	0.000	0.60	0.000	**0.56**	+0.10	w-p
Usha	work-relevant									0.35	0.002	0.22	0.061	**0.28**	-0.17	w-r

Model label:	V-M1-A	V-M1-B	V-M2	B-M2	V4-M2	V4o-M2
Percentage of significant rows:	24%	13%	9%	24%	82%	98%
Mean rpb (significant only, excl. ALL):	**0.19**	**0.17**	**-0.14**	**0.25**	**0.40**	**0.48**
Mean rpb (significant only, ALL only):	**0.13**	**0.04**	**n/a**	**0.12**	**0.36**	**0.49**

Table S11.5: r_{pb} for Evaluation Agreement by Persona, UD-ML Model and Evaluation Tag

Figure S11.6 shows the mean correlation (r_{pb}) between Modes 1 and 2 synthetic evaluation and participant classification actions for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.5.

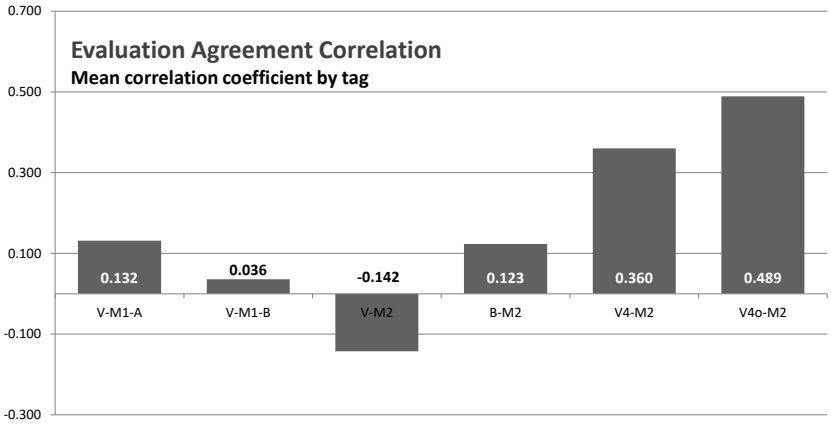


Figure S11.6: Mean Synthetic Evaluation *vs* Participant Classification r_{pb} by Eval Tag

Figure S11.7 shows the percentages of rows in Table S11.5 for each evaluation tag having a significant correlation ($p \leq 0.06$). Only these row values are used to calculate the mean data that is illustrated in Figures S11.8.

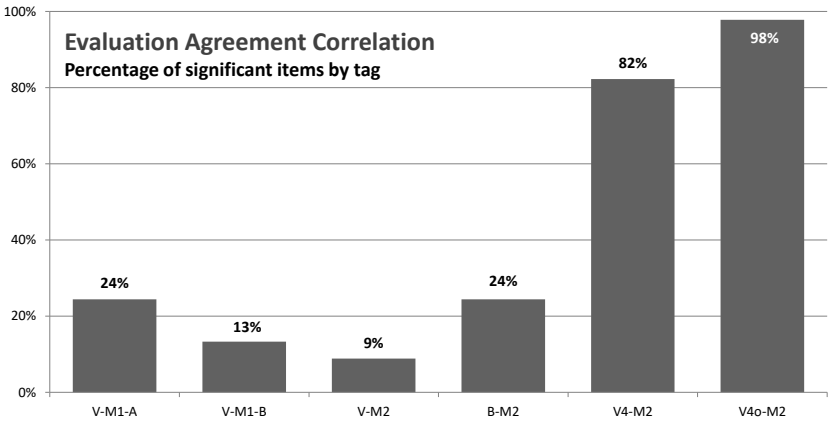
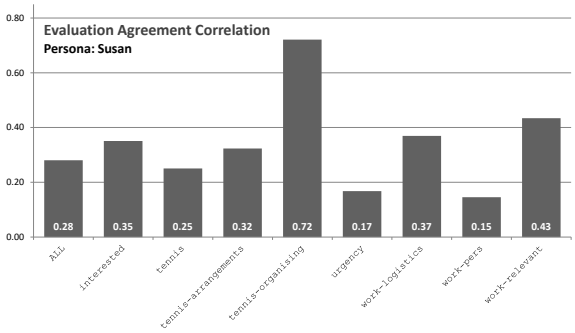
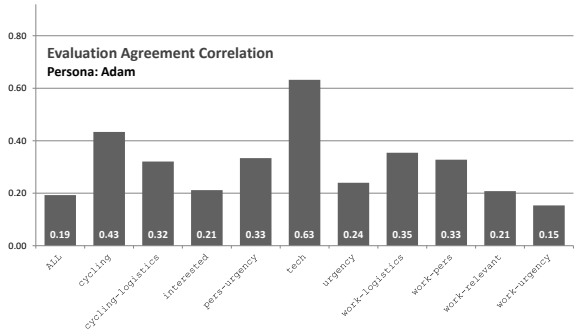


Figure S11.7: Percentage of Items Having Significant Correlation (r_{pb}) by Evaluation Tag

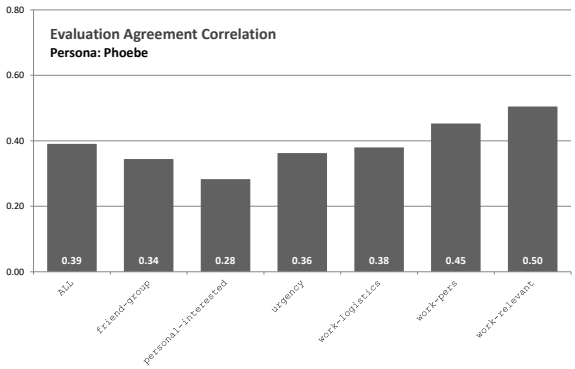
Figures S11.8a to S11.8e show the mean correlation (r_{pb}) between Modes 1 and 2 synthetic evaluation and training classifications by UD-ML Model for each persona; this data is sourced from Table S11.5.



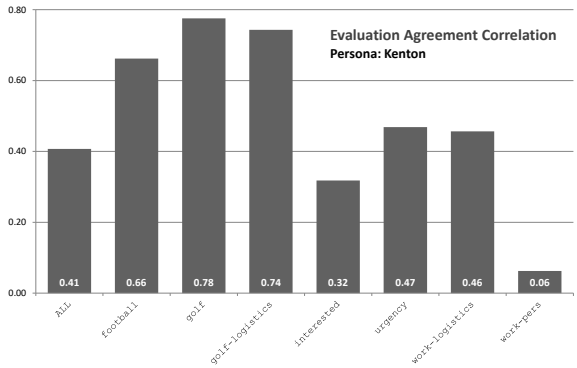
(a) Persona: Susan



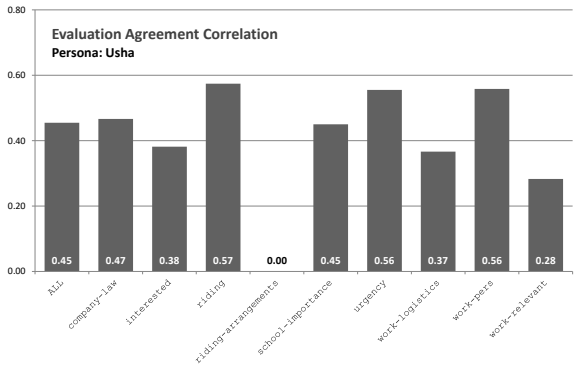
(b) Persona: Adam



(c) Persona: Phoebe



(d) Persona: Kenton



(e) Persona: Usha

Figure S11.8: Synthetic *vs* Participant Evaluation r_{pb} by UD-ML Model

S11.1.4 Evaluation Agreement (Mode 3)

Table S11.6 details the phi coefficient (r_ϕ) between Mode 3 synthetic evaluations and classification actions performed by the study participant during the training phase. In this case, r_ϕ is used because both Mode 3 evaluations and training classification actions are dichotomous (agree/disagree). The data is broken down by Persona, UD-ML Model and Evaluation Tag. Mean values of r_ϕ have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to r_ϕ. Only combinations of data having a significant correlation ($p \leq 0.06$) value are included in the mean.

Figure S11.9 shows the mean correlation (r_ϕ) between Mode 3 synthetic evaluation and participant classification actions for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.6.

Figure S11.10 shows the percentages of rows in Table S11.6 for each evaluation tag having a significant correlation ($p \leq 0.06$). Only these row values are used to calculate the mean data that is illustrated in Figures S11.11.

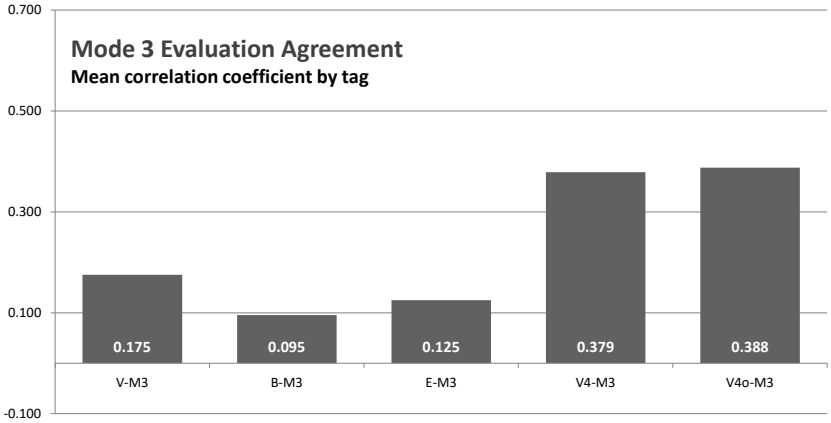


Figure S11.9: Mean Synthetic Evaluation *vs* Participant Classification r_ϕ by Eval Tag

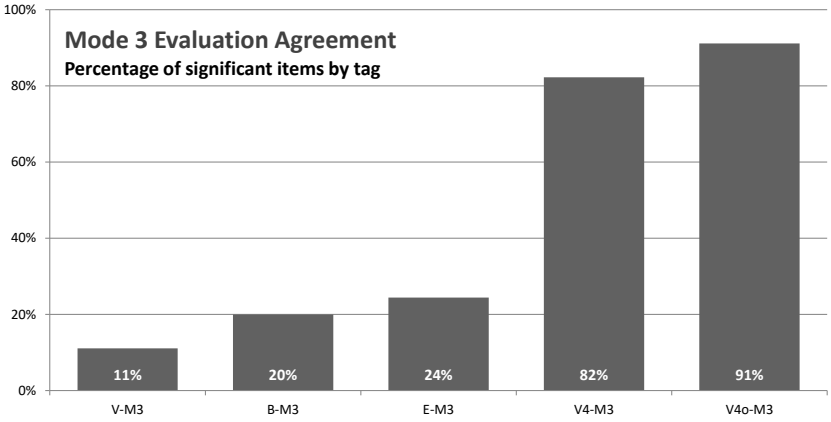


Figure S11.10: Percentage of Items Having Significant Correlation (r_ϕ) by Evaluation Tag

Phi Coefficient For Mode3 Agreement

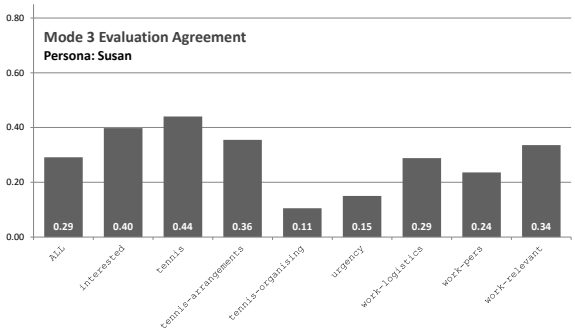
This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the phi coefficient between Mode 3 synthetic evaluations and classification actions entered by the participant. The higher the coeffieient, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those phi values having a significant correlation ($p < 0.06$); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 3 evaluations have dichotomous values, and this data only includes items for which the participant entered classification values during the training phase of the study.

Persona	Model	vanilla-mode3-01		base-mode3-01		ext-mode3-01		vanilla4-mode3-01		vanilla4o-mode3-01		Mean r		
		phi	*p*	*phi*	*p*	*phi*	*p*	*phi*	*p*	*phi*	*p*			
Susan	ALL	0.01	0.583	0.10	0.000	0.16	0.000	0.44	0.000	0.44	0.000	**0.29**		ALL
Susan	interested	0.04	0.495	0.07	0.282	0.03	0.680	0.46	0.000	0.33	0.000	**0.40**	+0.11	in
Susan	tennis	0.02	0.762	0.00	1.000	0.14	0.044	0.47	0.000	0.65	0.000	**0.44**	+0.15	tn
Susan	tennis-arrangements	0.37	0.000	0.04	0.537	0.09	0.152	0.24	0.004	0.45	0.000	**0.36**	+0.06	t-a
Susan	tennis-organising	0.01	0.829	0.00	1.000	0.00	0.950	0.12	0.149	0.20	0.003	**0.11**	-0.19	t-o
Susan	urgency	0.11	0.056	0.19	0.005	0.00	1.000	0.13	0.111	0.00	1.000	**0.15**	-0.14	ur
Susan	work-logistics	0.15	0.013	0.10	0.137	0.11	0.102	0.38	0.000	0.33	0.000	**0.29**	-0.00	w-l
Susan	work-pers	0.03	0.670	0.10	0.158	0.07	0.331	0.34	0.000	0.30	0.000	**0.24**	-0.06	w-p
Susan	work-relevant	0.12	0.045	0.14	0.036	0.18	0.009	0.61	0.000	0.41	0.000	**0.34**	+0.04	w-r
Adam	ALL	0.01	0.635	0.08	0.000	0.09	0.000	0.26	0.000	0.30	0.000	**0.18**		ALL
Adam	cycling	0.00	1.000	0.07	0.211	0.15	0.016	0.20	0.001	0.45	0.000	**0.27**	+0.09	cy
Adam	cycling-logistics	0.00	1.000	0.21	0.008	0.07	0.418	0.46	0.000	0.54	0.000	**0.41**	+0.23	c-l
Adam	interested	0.04	0.430	0.00	1.000	0.04	0.550	0.23	0.000	0.14	0.001	**0.18**	+0.00	in
Adam	pers-urgency	0.01	0.914	0.04	0.424	0.09	0.144	0.08	0.202	0.31	0.000	**0.31**	+0.13	p-u
Adam	tech	0.00	1.000	0.12	0.021	0.00	1.000	0.52	0.000	0.47	0.000	**0.38**	+0.20	tc
Adam	urgency	0.04	0.444	0.00	1.000	0.11	0.062	0.04	0.542	0.19	0.000	**0.15**	-0.03	ur
Adam	work-logistics	0.00	1.000	0.01	0.889	0.00	1.000	0.26	0.000	0.44	0.000	**0.35**	+0.17	w-l
Adam	work-pers	0.07	0.194	0.07	0.216	0.04	0.624	0.30	0.000	0.23	0.000	**0.27**	+0.08	w-p
Adam	work-relevant	0.12	0.018	0.21	0.000	0.11	0.053	0.14	0.019	0.28	0.000	**0.18**	-0.01	w-r
Adam	work-urgency	0.01	0.815	0.05	0.325	0.06	0.311	0.24	0.000	0.29	0.000	**0.26**	+0.08	w-u
Phoebe	ALL							0.35	0.000	0.35	0.000	**0.35**		ALL
Phoebe	friend-group							0.29	0.000	0.37	0.000	**0.33**	-0.02	f-g
Phoebe	personal-interested							0.24	0.000	0.21	0.001	**0.22**	-0.13	p-i
Phoebe	urgency							0.22	0.001	0.41	0.000	**0.32**	-0.04	ur
Phoebe	work-logistics							0.28	0.000	0.39	0.000	**0.33**	-0.02	w-l
Phoebe	work-pers							0.37	0.000	0.22	0.001	**0.30**	-0.06	w-p
Phoebe	work-relevant							0.59	0.000	0.55	0.000	**0.57**	+0.21	w-r
Kenton	ALL	0.07	0.084	0.11	0.002	0.13	0.001	0.47	0.000	0.41	0.000	**0.29**		ALL
Kenton	football	0.02	0.864	0.00	1.000	0.05	0.609	0.71	0.000	0.75	0.000	**0.73**	+0.44	fo
Kenton	golf	0.00	1.000	0.02	0.864	0.23	0.017	0.27	0.014	0.28	0.005	**0.26**	-0.03	go
Kenton	golf-logistics	0.02	0.828	0.26	0.006	0.44	0.000	0.74	0.000	0.71	0.000	**0.57**	+0.28	g-l
Kenton	interested	0.13	0.250	0.02	0.812	0.18	0.076	0.25	0.028	0.15	0.124	**0.25**	-0.03	in
Kenton	urgency	0.00	1.000	0.00	1.000	0.00	1.000	0.13	0.278	0.10	0.318	**0.10**	-0.19	ur
Kenton	work-logistics	0.00	1.000	0.00	1.000	0.02	0.854	0.18	0.122	0.31	0.002	**0.00**	-0.29	w-l
Kenton	work-pers	0.05	0.671	0.00	1.000	0.05	0.590	0.41	0.000	0.28	0.005	**0.35**	+0.06	w-p
Usha	ALL							0.36	0.000	0.43	0.000	**0.40**		ALL
Usha	company-law							0.41	0.000	0.41	0.000	**0.41**	+0.01	c-l
Usha	interested							0.30	0.006	0.35	0.001	**0.32**	-0.08	in
Usha	riding							0.25	0.020	0.46	0.000	**0.36**	-0.04	ri
Usha	riding-arrangements							0.08	0.429	0.34	0.001	**0.34**	-0.05	r-a
Usha	school-importance							0.20	0.065	0.32	0.002	**0.32**	-0.07	s-i
Usha	urgency							0.10	0.368	0.15	0.151	**0.10**	-0.30	ur
Usha	work-logistics							0.20	0.061	0.13	0.228	**0.20**	-0.20	w-l
Usha	work-pers							0.35	0.001	0.48	0.000	**0.42**	+0.02	w-p
Usha	work-relevant							0.27	0.011	0.31	0.003	**0.29**	-0.10	w-r

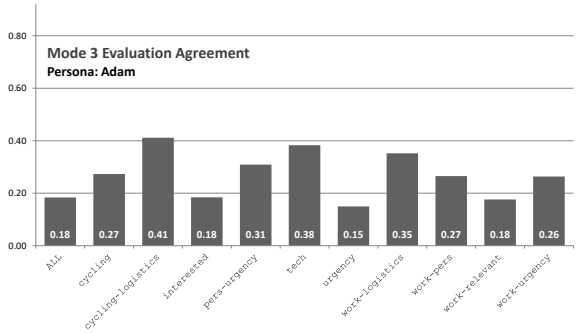
Model label:	V-M3	B-M3	E-M3	V4-M3	V4o-M3
Percentage of significant rows:	11%	20%	24%	82%	91%
Mean phi (significant only, excl. ALL):	**0.17**	**0.19**	**0.16**	**0.35**	**0.37**
Mean phi (significant only, ALL only):	**n/a**	**0.10**	**0.13**	**0.38**	**0.39**

Table S11.6: r_ϕ for Mode 3 Evaluation Agreement by Persona, UD-ML Model and Evaluation Tag

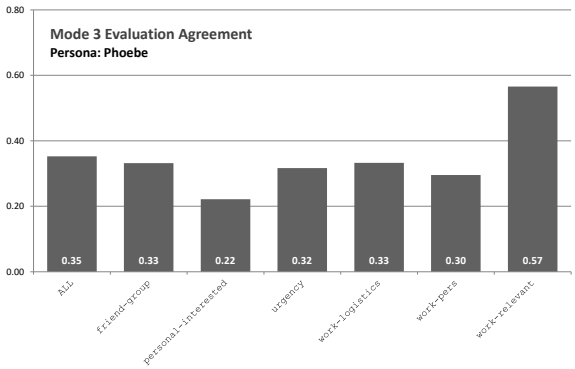
Figures S11.11a to S11.11e show the mean coefficient (r_ϕ) between Mode 3 synthetic evaluation and participant training classifications by UD-ML Model for each persona; this data is sourced from Table S11.6.



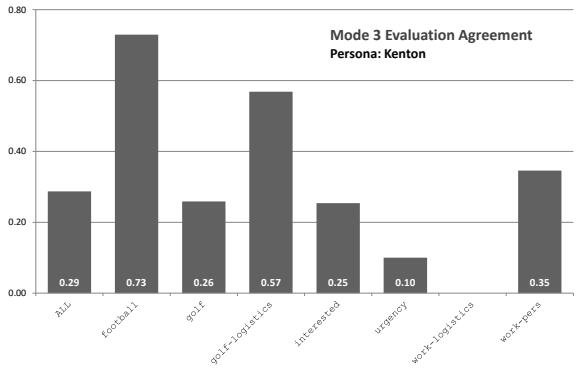
(a) Persona: Susan



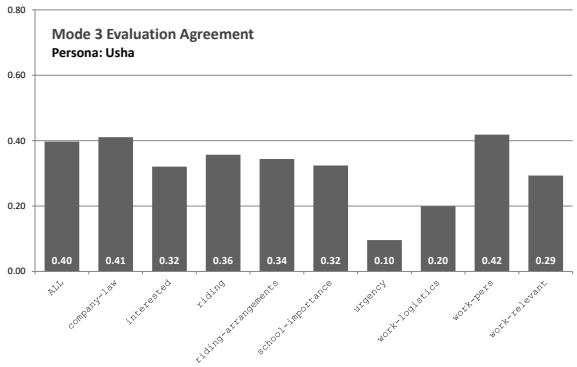
(b) Persona: Adam



(c) Persona: Phoebe



(d) Persona: Kenton



(e) Persona: Usha

Figure S11.11: Mode 3 Synthetic Evaluation *vs* Participant Classification r_ϕ by UD-ML Model

S11.1.5 Evaluation Difference

Table S11.7 shows aggregated data on the differences between numerical Likert values for synthetic evaluations and the corresponding participant feedback.

Figures S11.12 and S11.13 show this data broken down by evaluation output tag and UD-ML model respectively.

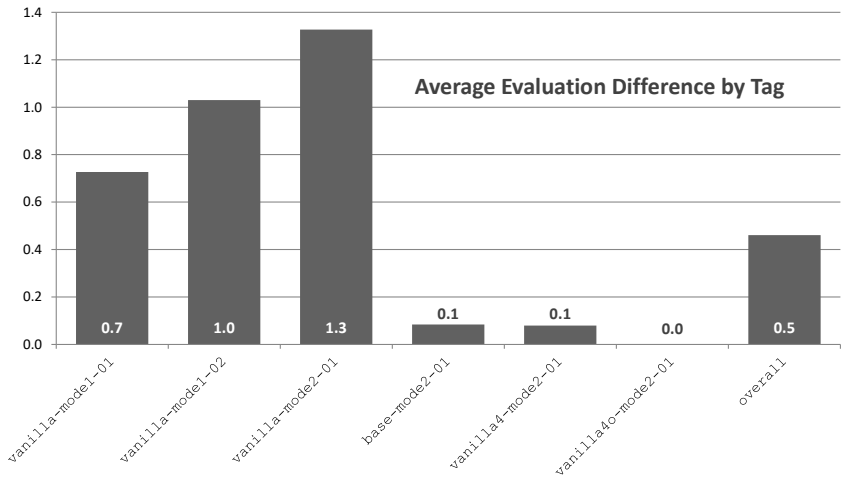


Figure S11.12: Average Evaluation Difference by Evaluation Tag

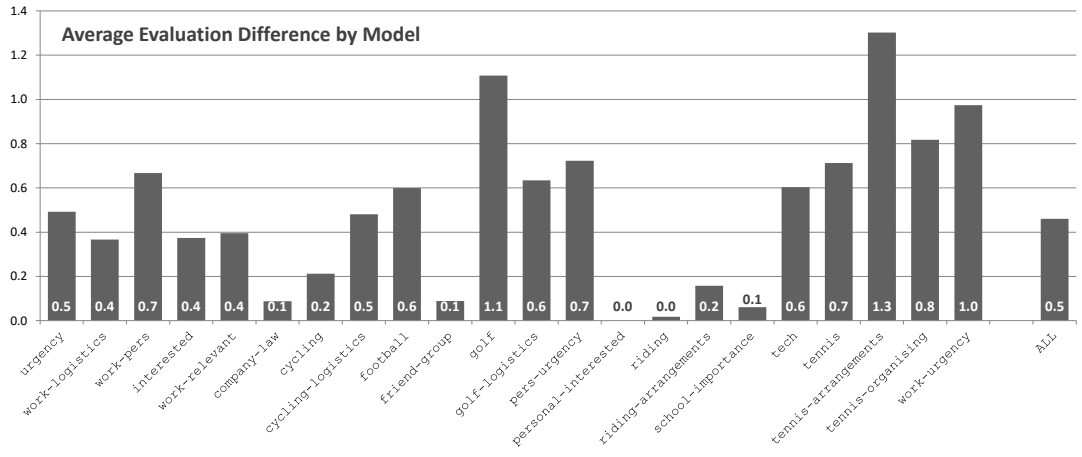


Figure S11.13: Average Evaluation Difference by UD-ML Model

Average and Spread for Evaluation Difference

This table shows the aggregated numerical differences between synthetic evaluations and the corresponding participant feedback actions, based on Likert value range of 1-5. A different close to zero indicates near perfect agreement on average, while a larger number represents a greater disagreement. The maximum value is 4, representing the difference between likert values 1 & 5 or 5 & 1.

		vanilla-mode1-01			vanilla-mode1-02			vanilla-mode2-01			base-mode2-01			vanilla4-mode2-01			vanilla4o-mode2-01			overall			
Model	**#**	**average**	**stddev**	**count**	**average**	**stddev**	**count**	**average**	**stddev**	**count**	**average**	**stddev**	**count**	**average**	**stddev**	**count**	**average**	**stddev**	**count**	**average**		**stddev**	
urgency	5	0.7	1.4	114	1.1	1.5	234	0.9	1.7	106	0.2	0.8	98	0.1	0.8	152	0.1	0.5	272	**0.5**	+0.0	**1.2**	ur
work-logistics	5	0.5	1.5	114	0.5	1.6	232	1.3	1.7	104	0.4	1.3	96	0.1	0.8	153	0.0	0.5	272	**0.4**	-0.1	**1.2**	w-l
work-pers	5	0.7	1.9	114	0.9	1.9	232	1.3	1.9	105	0.9	1.9	88	0.4	1.1	153	0.4	0.9	271	**0.7**	+0.2	**1.6**	w-p
interested	4	0.4	1.5	114	0.5	1.8	233	0.9	1.6	106	0.5	1.8	97	0.2	1.1	133	0.1	0.8	247	**0.4**	-0.1	**1.5**	in
work-relevant	4	0.4	2.0	73	0.6	1.9	117	1.3	2.1	69	0.1	1.4	64	0.2	0.8	130	0.2	0.7	197	**0.4**	-0.1	**1.4**	w-r
company-law	1													0.1	1.0	43	0.1	0.3	58	**0.1**	-0.4	**0.7**	co-l
cycling	1	0.1	1.2	30	0.1	0.8	29	0.4	1.3	37	0.3	1.3	36	0.2	0.9	42	0.1	1.2	22	**0.2**	-0.2	**1.1**	cy
cycling-logistics	1	0.1	0.5	15	0.2	1.1	17	1.4	1.2	20	0.5	1.5	22	0.0	1.1	26	0.4	1.3	18	**0.5**	+0.0	**1.2**	cy-l
football	1	0.7	1.3	41	1.0	1.8	115	1.5	1.4	37	0.0	1.0	33	0.0	0.0	23	0.0	0.7	75	**0.6**	+0.1	**1.3**	fo
friend-group	1													0.1	1.3	20	0.1	0.5	25	**0.1**	-0.4	**1.0**	f-g
golf	1	2.1	1.3	42	1.9	1.6	121	2.1	1.0	37	0.2	1.1	34	0.1	0.6	23	0.0	0.9	75	**1.1**	+0.6	**1.2**	go
golf-logistics	1	0.9	1.7	41	1.0	1.7	116	1.7	1.2	37	0.0	0.9	34	0.0	0.2	23	0.0	0.9	75	**0.6**	+0.2	**1.3**	g-l
pers-urgency	1	0.5	1.2	30	1.3	1.9	29	1.5	1.6	37	0.4	1.1	36	0.1	1.0	41	0.2	0.7	22	**0.7**	+0.3	**1.3**	p-u
personal-interested	1													0.1	0.7	20	0.0	0.4	25	**0.0**	-0.5	**0.6**	p-i
riding	1													0.1	0.5	43	0.0	0.7	58	**0.0**	-0.4	**0.6**	ri
riding-arrangements	1													0.2	0.7	43	0.1	0.4	58	**0.2**	-0.3	**0.5**	r-a
school-importance	1													0.1	1.0	43	0.1	0.6	58	**0.1**	-0.4	**0.8**	s-i
tech	1	0.2	1.1	30	1.2	1.8	29	1.7	1.6	37	0.0	0.6	36	0.1	0.3	41	0.1	0.2	22	**0.6**	+0.1	**1.1**	tc
tennis	1	1.6	1.5	43	1.4	1.9	88	1.3	2.1	32	0.4	1.8	30	0.0	1.6	25	0.1	0.6	92	**0.7**	+0.3	**1.6**	tn
tennis-arrangements	1	1.8	1.3	21	2.6	1.7	94	2.3	1.4	32	0.9	2.1	30	0.0	1.5	26	0.1	0.6	92	**1.3**	+0.8	**1.4**	t-a
tennis-organising	1	1.7	1.8	43	1.7	2.1	88	1.4	1.8	32	0.1	1.0	28	0.5	1.4	26	0.1	0.8	92	**0.8**	+0.4	**1.6**	t-o
work-urgency	1	0.8	1.3	30	2.0	1.7	30	2.2	1.6	37	0.3	1.1	37	0.2	0.6	41	0.1	0.4	22	**1.0**	+0.5	**1.2**	w-u
ALL		**0.7**	**1.6**	**895**	**1.0**	**1.8**	**1,804**	**1.3**	**1.8**	**865**	**0.1**	**1.5**	**799**	**0.1**	**1.0**	**1,270**	**0.0**	**0.7**	**2,148**	**0.5**		**1.4**	ALL
Model label:		V-M1-A			V-M1-B			V-M2			B-M2			V4-M2			V4o-M2						

Table S11.7: Average and Spread for Evaluation Difference

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

S11.1.6 Evaluation Ratings

Table S11.8 shows overall ratings values for Synthetic Evaluations, identified by Output Tag[1]. These values are calculated by taking the Evaluation Feedback [S11.1.2], Mode 1/2 Evaluation Agreement [S11.1.3] and Mode 3 Evaluation Agreement [S11.1.4] as applicable to the mode. In the case of Modes 1 and 2, where we have two out of these available, the overall rating is a mean of the values. In the case of Mode 3, we use the single value that is available.

Note that these comparisons are not all like-for-like – while all of these source statistics measure correlation between synthetic evaluation results and some yardstick (either participant feedback or comparison with phase 2 classification actions), they are calculated in different ways (r, r_{pb} and r_{ϕ} respectively). However, they provide a relative comparison of the strengths of each tag.

This data is also summarised in Figure S11.14.

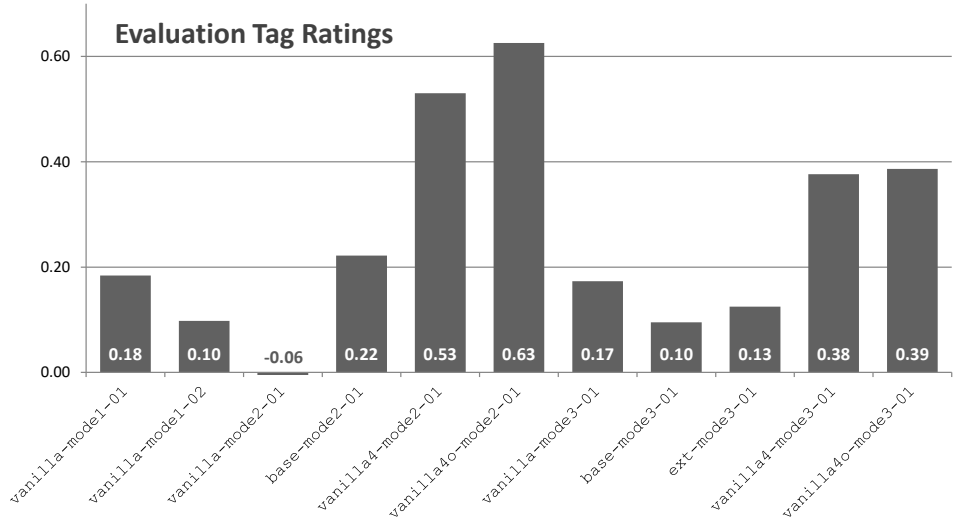


Figure S11.14: Synthetic Evaluation Tag Ratings

[1] Which represents a unique combination of OpenAI model, evaluation mode and iteration for the synthetic evaluation

Evaluation Tag Ratings

Rating values for models, taken from the mean of the 'ALL' correlation number across all personas using different techniques as applicable to the evaluation type. Where an ALL mean is not available, a mean of the individual items is used.

Tag	PFEF	PFEA	PCFMA	Overall
vanilla-mode1-01	0.24	0.13		0.18
vanilla-mode1-02	0.16	0.04		0.10
vanilla-mode2-01	0.01	-0.14		-0.06
base-mode2-01	0.32	0.12		0.22
vanilla4-mode2-01	0.70	0.36		0.53
vanilla4o-mode2-01	0.76	0.49		0.63
vanilla-mode3-01			0.17	0.17
base-mode3-01			0.10	0.10
ext-mode3-01			0.13	0.13
vanilla4-mode3-01			0.38	0.38
vanilla4o-mode3-01			0.39	0.39

PFEF: Pearson Correlation for Evaluation Feedback
PFEA: Point Biserial Correlation for Evaluation Agreement
PCFMA: Phi Coefficient for Mode3 Evaluation Agreement

Table S11.8: Synthetic Evaluation Tag Ratings

S11.1.7 Evaluation Results

Table S11.9 shows detailed results of Synthetic Evaluations of the performance of each UD-ML model. Rows represent the UD-ML models under evaluation, and columns represent the Output Tags of the evaluations. The evaluations are output as two metrics:

- **Evaluation Percentage**
  - For Mode 1&2 evaluations, this is defined as the percentage of evaluated items awarded Likert values Agree & Strongly Agree)
  - For Mode 3 evaluations, this is defined as the percentage of evaluated items that have an Agree result
- **Evaluation Score**
  - For Mode 1&2 evaluations, a score is assigned to items based on the Evaluation Likert value:
    - Stongly Disagree: 0
    - Disagree: 0.25
    - Neutral: 0.5
    - Agree: 0.75
    - Stongly Agree: 1
  - For Mode 3 evaluations, an award of Agree is assigned a score of 1 and Disagree is assigned 0
  - A mean is then calculated for these scores

Table S11.10 shows the Evaluation scores from Table S11.9 consolidated for UD-ML model via a simple mean and compared to the Manual Classification Agreement values taken from Tables S11.1 and S11.3. We have also added two metrics to measure similarity between the values, *Proximity Ratio* and *Squared Error x100* (SE_{100}), defined as:

$$\text{Proximity Ratio} = \frac{\min(A, B)}{\max(A, B)}$$

$$\text{SE}_{100} = 100 \times (A - B)^2$$

Where A is the value for Percentage of Manual Classification Agreement [S11.1], and B is the Evaluation Score for the `V4o-M2` column (corresponding to the `vanilla4o-mode2-01` tag). This tag is selected as the prime Synthetic Evaluation as it has the highest Evaluation Rating as defined in Section S11.1.6.

Table S11.11 shows a subset of this information - containing only the agreement percentage, evaluation score and similarity metrics for clarity. This is shown in graphical form in Figure S11.15

Evaluation Results

Persona	Model	vanilla-mode1-01		vanilla-mode1-02		vanilla-mode2-01		base-mode2-01		vanilla4-mode2-01		vanilla4o-mode2-01		vanilla-mode3-01		base-mode3-01		ext-mode3-01		vanilla4-mode3-01		vanilla4o-mode3-01		All Models		Best Model		
		agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	
Susan	ALL	59.9%	0.70	49.2%	0.57	35.6%	0.49	79.7%	0.83	76.1%	0.79	85.8%	0.87	40.9%	0.41	54.2%	0.54	45.0%	0.45	79.4%	0.79	82.2%	0.82	**62.6%**	**0.66**	**85.8%**	**0.87**	ALL
Susan	interested	39.9%	0.55	45.8%	0.55	32.1%	0.48	53.1%	0.59	48.2%	0.57	62.7%	0.67	42.6%	0.43	34.7%	0.35	25.4%	0.25	65.6%	0.66	60.2%	0.60	**46.4%**	**0.52**	**62.7%**	**0.67**	in
Susan	tennis	55.5%	0.67	35.5%	0.45	35.3%	0.47	91.0%	0.92	81.7%	0.84	92.0%	0.93	33.5%	0.33	34.7%	0.35	33.4%	0.33	84.4%	0.84	88.3%	0.88	**60.5%**	**0.64**	**92.0%**	**0.93**	tn
Susan	tennis-arrangements	12.0%	0.49	1.8%	0.22	3.8%	0.23	58.6%	0.65	83.6%	0.85	96.5%	0.96	4.1%	0.04	33.5%	0.33	50.0%	0.50	84.1%	0.84	92.8%	0.93	**47.3%**	**0.55**	**96.5%**	**0.96**	t-a
Susan	tennis-organising	45.3%	0.62	27.8%	0.41	35.1%	0.48	91.0%	0.92	82.7%	0.83	96.5%	0.96	38.3%	0.38	50.6%	0.51	51.1%	0.51	82.7%	0.83	91.0%	0.91	**62.9%**	**0.67**	**96.5%**	**0.96**	t-o
Susan	urgency	81.9%	0.80	73.0%	0.74	51.0%	0.65	95.8%	0.96	87.7%	0.89	98.8%	0.96	74.6%	0.75	88.3%	0.88	67.2%	0.67	91.7%	0.92	98.8%	0.99	**82.6%**	**0.84**	**98.8%**	**0.96**	ur
Susan	work-logistics	87.4%	0.88	75.4%	0.78	48.5%	0.59	96.7%	0.96	87.2%	0.89	92.0%	0.93	53.2%	0.53	70.0%	0.70	49.9%	0.50	86.5%	0.87	87.9%	0.88	**75.9%**	**0.77**	**92.0%**	**0.93**	w-l
Susan	work-pers	68.0%	0.74	66.6%	0.71	35.0%	0.47	60.3%	0.66	54.4%	0.59	55.3%	0.62	37.1%	0.37	45.8%	0.46	13.0%	0.13	58.3%	0.58	51.6%	0.52	**49.6%**	**0.53**	**55.3%**	**0.62**	w-p
Susan	work-relevant	88.4%	0.88	72.8%	0.76	43.9%	0.55	90.7%	0.92	83.6%	0.84	92.9%	0.92	43.7%	0.44	75.3%	0.75	69.3%	0.69	82.0%	0.82	87.3%	0.87	**75.4%**	**0.77**	**92.9%**	**0.92**	w-r
Adam	ALL	73.7%	0.79	57.0%	0.65	32.5%	0.51	84.6%	0.86	83.2%	0.85	89.4%	0.89	54.0%	0.54	64.1%	0.64	75.8%	0.76	84.5%	0.85	85.2%	0.85	**71.3%**	**0.74**	**89.4%**	**0.89**	ALL
Adam	cycling	94.8%	0.95	90.7%	0.92	70.6%	0.82	91.2%	0.91	92.8%	0.93	92.7%	0.93	74.2%	0.74	70.6%	0.71	75.1%	0.75	92.9%	0.93	90.9%	0.91	**85.1%**	**0.86**	**92.7%**	**0.93**	cy
Adam	cycling-logistics	91.8%	0.93	90.8%	0.91	49.5%	0.67	88.6%	0.93	86.7%	0.88	89.8%	0.89	84.0%	0.84	77.2%	0.77	85.5%	0.85	78.3%	0.78	89.2%	0.89	**82.9%**	**0.85**	**89.8%**	**0.89**	c-l
Adam	interested	68.6%	0.74	65.2%	0.71	39.4%	0.53	60.8%	0.66	63.3%	0.69	74.0%	0.76	46.4%	0.46	42.6%	0.43	40.6%	0.41	67.7%	0.68	66.2%	0.66	**57.7%**	**0.61**	**74.0%**	**0.76**	in
Adam	pers-urgency	61.6%	0.69	42.8%	0.55	19.9%	0.44	92.2%	0.93	85.2%	0.87	92.2%	0.90	49.1%	0.49	76.6%	0.77	81.1%	0.81	88.9%	0.89	87.1%	0.87	**70.6%**	**0.75**	**92.2%**	**0.90**	p-u
Adam	tech	84.7%	0.88	62.0%	0.68	38.1%	0.59	94.2%	0.94	90.0%	0.90	94.3%	0.94	73.7%	0.74	82.6%	0.83	89.4%	0.89	87.7%	0.88	93.7%	0.94	**80.9%**	**0.84**	**94.3%**	**0.94**	tc
Adam	urgency	64.6%	0.74	41.9%	0.55	24.7%	0.47	95.1%	0.95	90.2%	0.92	95.3%	0.93	56.4%	0.56	76.0%	0.76	81.9%	0.82	92.5%	0.93	93.5%	0.93	**73.8%**	**0.78**	**95.3%**	**0.93**	ur
Adam	work-logistics	89.1%	0.90	68.3%	0.75	33.8%	0.52	95.1%	0.95	91.1%	0.91	95.1%	0.95	63.8%	0.64	66.8%	0.67	89.2%	0.89	91.0%	0.91	90.7%	0.91	**79.4%**	**0.82**	**95.1%**	**0.95**	w-l
Adam	work-pers	59.8%	0.69	56.0%	0.63	25.6%	0.38	62.6%	0.66	64.1%	0.71	82.1%	0.82	31.3%	0.31	33.0%	0.33	40.1%	0.40	69.5%	0.70	69.4%	0.69	**53.9%**	**0.57**	**82.1%**	**0.82**	w-p
Adam	work-relevant	68.6%	0.76	46.2%	0.57	26.0%	0.45	80.6%	0.81	81.4%	0.84	86.5%	0.87	48.9%	0.49	54.9%	0.55	74.1%	0.74	82.2%	0.82	83.6%	0.84	**66.6%**	**0.70**	**86.5%**	**0.87**	w-r
Adam	work-urgency	63.6%	0.71	28.3%	0.44	7.4%	0.33	85.7%	0.87	88.7%	0.90	92.0%	0.92	29.1%	0.29	66.1%	0.66	86.0%	0.86	90.7%	0.91	90.3%	0.90	**66.2%**	**0.71**	**92.0%**	**0.92**	w-u
Phoebe	ALL									76.8%	0.80	84.3%	0.84							81.5%	0.82	81.9%	0.82	**81.1%**	**0.82**	**84.3%**	**0.84**	ALL
Phoebe	friend-group									82.2%	0.84	84.8%	0.87							82.4%	0.82	84.2%	0.84	**83.4%**	**0.84**	**84.8%**	**0.87**	f-g
Phoebe	personal-interested									62.8%	0.69	83.6%	0.79							76.8%	0.77	82.0%	0.82	**76.3%**	**0.77**	**83.6%**	**0.79**	p-i
Phoebe	urgency									79.1%	0.84	93.3%	0.92							87.7%	0.88	92.0%	0.92	**88.0%**	**0.89**	**93.3%**	**0.92**	ur
Phoebe	work-logistics									77.8%	0.80	86.0%	0.88							81.7%	0.82	85.3%	0.85	**82.7%**	**0.84**	**86.0%**	**0.88**	w-l
Phoebe	work-pers									79.7%	0.81	73.9%	0.77							80.0%	0.80	64.4%	0.64	**74.5%**	**0.76**	**73.9%**	**0.77**	w-p
Phoebe	work-relevant									79.4%	0.81	83.9%	0.84							80.7%	0.81	83.1%	0.83	**81.8%**	**0.82**	**83.9%**	**0.84**	w-r
Kenton	ALL	57.6%	0.67	49.7%	0.60	29.3%	0.53	79.7%	0.82	82.1%	0.85	84.3%	0.86	60.3%	0.60	69.8%	0.70	82.3%	0.82	84.4%	0.84	81.3%	0.81	**69.2%**	**0.74**	**84.3%**	**0.86**	ALL
Kenton	football	73.7%	0.79	59.2%	0.65	43.7%	0.61	85.8%	0.86	92.6%	0.93	88.4%	0.89	60.5%	0.60	60.8%	0.61	83.3%	0.83	91.2%	0.91	89.4%	0.89	**75.3%**	**0.78**	**88.4%**	**0.89**	fo
Kenton	golf	20.0%	0.46	15.8%	0.38	11.7%	0.43	82.7%	0.85	96.0%	0.96	97.8%	0.98	46.8%	0.47	79.4%	0.79	89.5%	0.90	96.2%	0.96	97.3%	0.97	**66.7%**	**0.74**	**97.8%**	**0.98**	go
Kenton	golf-logistics	74.9%	0.81	61.7%	0.68	30.6%	0.56	95.7%	0.96	96.9%	0.97	97.4%	0.98	74.1%	0.74	91.1%	0.91	90.7%	0.91	96.2%	0.96	96.8%	0.97	**82.4%**	**0.86**	**97.4%**	**0.98**	gl
Kenton	interested	49.8%	0.62	49.2%	0.59	16.3%	0.44	53.2%	0.58	55.5%	0.66	62.5%	0.68	49.5%	0.50	42.9%	0.43	77.3%	0.77	62.3%	0.62	51.0%	0.51	**51.8%**	**0.58**	**62.5%**	**0.68**	in
Kenton	urgency	54.4%	0.63	45.1%	0.57	31.3%	0.59	92.7%	0.94	79.6%	0.83	94.5%	0.91	86.7%	0.87	91.3%	0.91	93.2%	0.93	82.5%	0.83	93.5%	0.93	**76.8%**	**0.81**	**94.5%**	**0.91**	ur
Kenton	work-logistics	84.6%	0.85	71.3%	0.76	45.3%	0.66	98.3%	0.98	87.0%	0.88	83.5%	0.86	76.1%	0.76	83.4%	0.83	91.4%	0.91	89.0%	0.89	79.9%	0.80	**80.9%**	**0.83**	**83.5%**	**0.86**	w-l
Kenton	work-pers	46.8%	0.56	47.3%	0.55	26.4%	0.40	46.7%	0.53	67.4%	0.71	66.3%	0.71	28.5%	0.28	37.8%	0.38	49.0%	0.49	73.3%	0.73	61.2%	0.61	**50.1%**	**0.54**	**66.3%**	**0.71**	w-p
Usha	ALL									85.7%	0.87	89.4%	0.89							87.4%	0.87	86.4%	0.86	**87.2%**	**0.87**	**89.4%**	**0.89**	ALL
Usha	company-law									93.6%	0.94	98.0%	0.97							91.8%	0.92	92.6%	0.93	**94.0%**	**0.94**	**98.0%**	**0.97**	c-l
Usha	interested									70.4%	0.74	72.4%	0.72							77.9%	0.78	68.9%	0.69	**72.4%**	**0.73**	**72.4%**	**0.72**	in
Usha	riding									92.4%	0.92	90.5%	0.91							90.0%	0.90	91.0%	0.91	**91.0%**	**0.91**	**90.5%**	**0.91**	ri
Usha	riding-arrangements									87.2%	0.89	97.0%	0.96							88.3%	0.88	95.0%	0.95	**91.9%**	**0.92**	**97.0%**	**0.96**	r-a
Usha	school-importance									94.4%	0.95	93.5%	0.94							96.8%	0.97	95.3%	0.95	**95.0%**	**0.95**	**93.5%**	**0.94**	s-i
Usha	urgency									88.4%	0.90	96.5%	0.93							90.4%	0.90	93.5%	0.93	**92.2%**	**0.92**	**96.5%**	**0.93**	ur
Usha	work-logistics									88.8%	0.90	90.5%	0.89							87.2%	0.87	83.5%	0.84	**87.5%**	**0.88**	**90.5%**	**0.89**	w-l
Usha	work-pers									65.6%	0.68	72.9%	0.75							73.2%	0.73	69.8%	0.70	**70.4%**	**0.72**	**72.9%**	**0.75**	w-p
Usha	work-relevant									90.4%	0.91	93.5%	0.93							91.1%	0.91	87.9%	0.88	**90.7%**	**0.91**	**93.5%**	**0.93**	w-r
Mean		**65.0%**	**0.73**	**53.4%**	**0.62**	**32.9%**	**0.51**	**81.5%**	**0.84**	**81.2%**	**0.83**	**87.0%**	**0.87**	**52.2%**	**0.52**	**62.6%**	**0.63**	**67.1%**	**0.67**	**83.6%**	**0.84**	**83.7%**	**0.84**	**74.3%**	**0.77**	**87.0%**	**0.87**	
Model label:		V-M1-A		V-M1-B		V-M2		B-M2		V4-M2		V4o-M2		V-M3		B-M3		E-M3		V4-M3		V4o-M3		All		Best		

Table S11.9: Evaluation Results Detail

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

Classification Manual Agreement vs Evaluation Score

This table shows the Classification Manual Agreement value of each UD-ML model alongside the Evaluation Score awarded by each of the Synthetic Evaluations. The figure in square brackets below the evaluation tag label is the rating for that tag, a composite coefficient that represents the quality of that evaluation tag; the closer this rating is to 1, the higher the quality. The final columns are a proximity ratio and squared error value, comparing the closeness of the best evaluation (V4o-M2) with the classification agreement.

		Manual		V-M1-A	V-M1-B	V-M2	B-M2	V4-M2	V4o-M2	V-M3	B-M3	E-M3	V4-M3	V4o-M3		Proximity	Squared
Model	**#**	**kappa**	**% agree**	**[0.18]**	**[0.10]**	**-[0.06]**	**[0.22]**	**[0.53]**	**[0.63]**	**[0.17]**	**[0.10]**	**[0.13]**	**[0.38]**	**[0.39]**		**Ratio**	**Error**
ALL		0.83	92.8%	0.72	0.61	0.51	0.84	0.83	0.87	0.52	0.63	0.68	0.83	0.83	ALL	0.94	0.33
urgency	5	0.71	97.0%	0.72	0.62	0.57	0.95	0.88	0.93	0.73	0.85	0.81	0.89	0.94	ur	0.96	0.16
work-logistics	5	0.64	96.7%	0.88	0.76	0.59	0.96	0.88	0.90	0.64	0.73	0.77	0.87	0.86	w-l	0.93	0.42
work-pers	5	0.75	91.0%	0.66	0.63	0.42	0.62	0.70	0.73	0.32	0.39	0.34	0.71	0.63	w-p	0.81	3.09
interested	4	0.71	85.6%	0.88	0.76	0.55	0.92	0.84	0.92	0.44	0.75	0.69	0.82	0.87	in	0.93	0.41
work-relevant	4	0.48	90.0%	0.82	0.67	0.50	0.87	0.85	0.89	0.47	0.65	0.72	0.84	0.86	w-r	0.99	0.01
company-law	1	0.61	94.1%	0.00	0.00	0.00	0.00	0.94	0.97	0.00	0.00	0.00	0.92	0.93	co-l	0.97	0.09
cycling	1	0.79	92.4%	0.95	0.92	0.82	0.91	0.93	0.93	0.74	0.71	0.75	0.93	0.91	cy	0.99	0.00
cycling-logistics	1	0.56	93.0%	0.93	0.91	0.67	0.93	0.88	0.89	0.84	0.77	0.85	0.78	0.89	cy-l	0.96	0.16
football	1	0.79	90.1%	0.79	0.65	0.61	0.86	0.93	0.89	0.60	0.61	0.83	0.91	0.89	fo	0.99	0.01
friend-group	1	0.74	89.0%	0.00	0.00	0.00	0.00	0.84	0.87	0.00	0.00	0.00	0.82	0.84	f-g	0.98	0.04
golf	1	0.87	98.6%	0.46	0.38	0.43	0.85	0.96	0.98	0.47	0.79	0.90	0.96	0.97	go	0.99	0.00
golf-logistics	1	0.55	95.7%	0.81	0.68	0.56	0.96	0.97	0.98	0.74	0.91	0.91	0.96	0.97	g-l	0.98	0.05
pers-urgency	1	0.75	94.7%	0.69	0.55	0.44	0.93	0.87	0.90	0.49	0.77	0.81	0.89	0.87	p-u	0.95	0.22
personal-interested	1	0.70	87.8%	0.00	0.00	0.00	0.00	0.69	0.79	0.00	0.00	0.00	0.77	0.82	p-i	0.90	0.77
riding	1	0.73	91.8%	0.00	0.00	0.00	0.00	0.92	0.91	0.00	0.00	0.00	0.90	0.91	ri	0.99	0.01
riding-arrangements	1	0.64	97.4%	0.00	0.00	0.00	0.00	0.89	0.96	0.00	0.00	0.00	0.88	0.95	r-a	0.99	0.02
school-importance	1	0.43	90.4%	0.00	0.00	0.00	0.00	0.95	0.94	0.00	0.00	0.00	0.97	0.95	s-i	0.96	0.13
tech	1	0.77	95.9%	0.88	0.68	0.59	0.94	0.90	0.94	0.74	0.83	0.89	0.88	0.94	tc	0.98	0.04
tennis	1	0.48	88.5%	0.67	0.45	0.47	0.92	0.84	0.93	0.33	0.35	0.33	0.84	0.88	tn	0.95	0.21
tennis-arrangements	1	0.51	97.4%	0.49	0.22	0.23	0.65	0.85	0.96	0.04	0.33	0.50	0.84	0.93	t-a	0.99	0.02
tennis-organising	1	0.31	96.6%	0.62	0.41	0.48	0.92	0.83	0.96	0.38	0.51	0.51	0.83	0.91	t-o	0.99	0.00
work-urgency	1	0.64	94.0%	0.71	0.44	0.33	0.87	0.90	0.92	0.29	0.66	0.86	0.91	0.90	w-u	0.98	0.04

Proximity Ratio Calculated as min(A,B)/max(A,B), where A is manual classification percentage agreement and B is V4o-M2 evaluation score

Squared Error Calculated as 100x(A-B)^2, where A is manual classification percentage agreement and B is V4o-M2 evaluation score - note that this is multiplied by 100 for readability

Table S11.10: Classification Agreement vs Evaluation Score for All Evaluations

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

Classification Manual Agreement vs Evaluation Score

This table shows the Classification Manual Agreement value of each UD-ML model alongside the V4o-M2 Synthetic Evaluation Score. A proximity ratio and squared error are calculated to show the closeness of the synthetic to manual classifications.

Model	#	Manual Agreement	Evaluation Score	Proximity Ratio	Squared Error
ALL		0.93	0.87	0.94	0.33
urgency	5	0.97	0.93	0.96	0.16
work-logistics	5	0.97	0.90	0.93	0.42
work-pers	5	0.91	0.73	0.81	3.09
interested	4	0.86	0.92	0.93	0.41
work-relevant	4	0.90	0.89	0.99	0.01
company-law	1	0.94	0.97	0.97	0.09
cycling	1	0.92	0.93	0.99	0.00
cycling-logistics	1	0.93	0.89	0.96	0.16
football	1	0.90	0.89	0.99	0.01
friend-group	1	0.89	0.87	0.98	0.04
golf	1	0.99	0.98	0.99	0.00
golf-logistics	1	0.96	0.98	0.98	0.05
pers-urgency	1	0.95	0.90	0.95	0.22
personal-interested	1	0.88	0.79	0.90	0.77
riding	1	0.92	0.91	0.99	0.01
riding-arrangements	1	0.97	0.96	0.99	0.02
school-importance	1	0.90	0.94	0.96	0.13
tech	1	0.96	0.94	0.98	0.04
tennis	1	0.88	0.93	0.95	0.21
tennis-arrangements	1	0.97	0.96	0.99	0.02
tennis-organising	1	0.97	0.96	0.99	0.00
work-urgency	1	0.94	0.92	0.98	0.04

Table S11.11: Classification Agreement vs Evaluation Score for Prime Evaluation

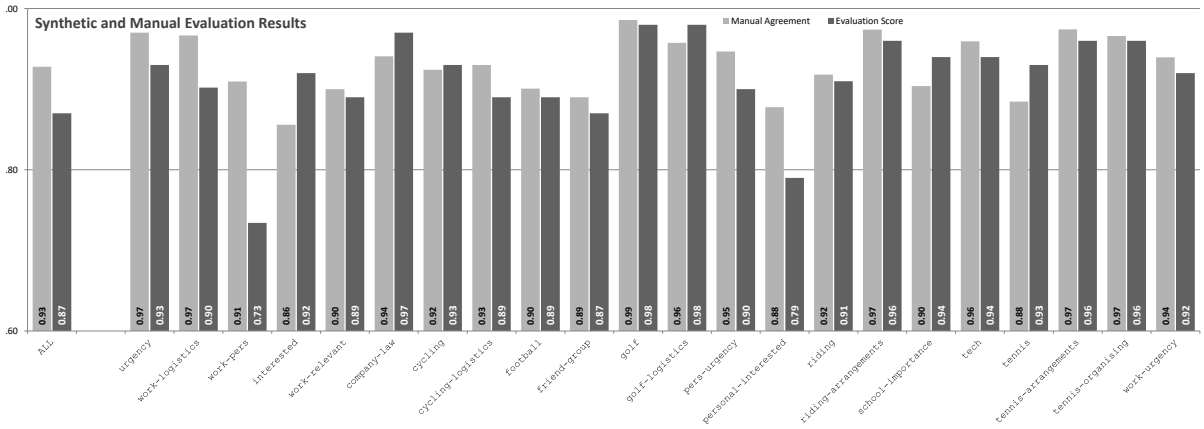


Figure S11.15: Classification Agreement vs Evaluation Score for Prime Evaluation

S11.2 Selection of Evaluation Items

This section contains a selection of individual synthetic evaluation item summaries, organised by Persona. Each item is formatted to show the type of item, channel that it was sent to, output tag, unique Item UUID and a summary of the item text[2]). Each entry shows what the classification (UD-ML model) was, the classification selected by UD-ML, the synthetic evaluation, and the participant-entered value. It also includes the evaluation text generated by the synthetic evaluator.

S11.2.1 Susan

S11.2.1.1 Susan Example 1 - vanilla - mode1

SLACK MESSAGE *family*
`phase2, phase2-vanilla-mode1-01`
`008c9eda-86bc-44f8-9a47-a36138577bbe`
If anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you w...

Classification: **interested**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Agree**

As Susan, I strongly disagree with the classification 'not' for this item. The message is related to my hobby of tennis, and I am interested in it from a personal perspective. The correct classification should have been 'personal'.

S11.2.1.3 Susan Example 2 - vanilla - mode1

SLACK MESSAGE *tennis-chat*
`phase2-vanilla-mode1-01`
`02f38c70-c331-48bb-8b09-3b25942ae1c8`
Definitely, Mason. It's the camaraderie off the court that makes us stronger on it.

Classification: **tennis**
Categorised as: **tennis**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I completely agree with the classification chosen by the system as the item appears to be related to the user's hobby/interest of tennis. It is a casual conversation related to tennis and seems to be specifically about the user's interest in tennis.

[2]Which is also a hyperlink to the item – in the case of Slack messages, this link will no longer be accessible, and we cannot guarantee that any other link type will remain valid

S11.2.1.5 Susan Example 3 - vanilla - mode1

SLACK MESSAGE *work-team-general*
phase2-vanilla-mode1-02
07f09430-3e9a-4a50-acef-1e66e448f288
Sophie, can you put me down as a maybe for Thursday? Waiting to hear back about a dentist appointmen...

Classification: **work-logistics**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Disagree**

The system correctly marked this as 'not' for work-logistics. The message is not related to work logistics, but rather about a personal dental appointment. My likert value is 5 as I completely agree with the system's classification.

S11.2.1.7 Susan Example 4 - vanilla - mode1

SLACK MESSAGE *susan-neil*
phase2-vanilla-mode1-02
165eec01-81fe-405e-9717-72df9a89b290
Good idea. Let's sit down this weekend and brainstorm some ideas.

Classification: **work-relevant**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

This message is not work-related and doesn't seem relevant to my work tasks at the university. The system's classification of 'not' is appropriate in this case.

S11.2.1.8 Susan Example 5 - vanilla - mode2

SLACK MESSAGE *family*
phase2-vanilla-mode2-01
008c9eda-86bc-44f8-9a47-a36138577bbe
If anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you w...

Classification: **work-relevant**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

The system classified this as 'not' work-relevant, which is correct as the content is about fresh eggs. However, the context of this message may be misinterpreted, so I am neutral about this classification.

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

S11.2.1.9 Susan Example 6 - vanilla - mode2

SLACK MESSAGE	*family*
phase2-vanilla-mode2-01	**@neil, can you check if we're out of bird seed? Noticed the feeder's looking a bit empty and the rob...**
031e5f68-da66-44c0-b056-abb01dd6be14	

Classification: **work-logistics**
Categorised as: **not**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Agree**

This message does not seem to be related to work logistics. It's about checking on bird seed and the robins, which is more personal or social. I mostly disagree with the 'work logistics' classification and would choose 'not' instead. Therefore, I have given an evaluationLikert value of 4.

S11.2.1.10 Susan Example 7 - vanilla4 - mode2

SLACK MESSAGE	*tennis-chat*
phase2-vanilla4-mode2-01	**Pressure's on then! But we've got this, team. Let's show them what we're made of! :muscle::tennis:**
0294b201-4cb2-4f2b-9182-9808cc83a9a4	

Classification: **interested**
Categorised as: **personal**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

Given my big interest in tennis, classifying this under 'personal' because it's about my hobby is perfectly aligned. Therefore, I strongly agree with this classification.

S11.2.1.11 Susan Example 8 - vanilla4 - mode2

SLACK MESSAGE	*tennis-chat*
phase2-vanilla4-mode2-01	**Pressure's on then! But we've got this, team. Let's show them what we're made of! :muscle::tennis:**
0294b201-4cb2-4f2b-9182-9808cc83a9a4	

Classification: **tennis-arrangements**
Categorised as: **not**
Synthetic Evaluation: **Disagree**
Participant Evaluation: **Disagree**

This message seems focused on team motivation and encouragement related to a tennis match, which leans more towards 'arrangements' rather than not being relevant at all. Therefore, I mostly disagree with the 'not' classification.

S11.2.1.12 Susan Example 9 - vanilla4 - mode2

RSS ITEM *UK news | The Guardian*
phase2-vanilla4-mode2-01 **Grocery price rises in Great Britain slow as cost of toilet rolls, butter and milk falls**
1d495126-cb4e-4a82-a7aa-2e438b504cc1

Classification: **work-relevant**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification 'not' for relevance to my work tasks. The content is focused on grocery market trends, which is unrelated to my administrative duties at the university.

S11.2.1.13 Susan Example 10 - vanilla4o - mode2

SLACK MESSAGE *work-team-general*
phase2-vanilla4o-mode2-01 **Eleanor, thanks for the heads up! Will do. And Alisha, I thought it was just me! Glad to know I'm no...**
01df244a-719c-4c51-a086-4b2dd2b7407f

Classification: **interested**
Categorised as: **not**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Agree**

I mostly agree with the classification of 'not'. There isn't explicit content in the message that relates to any of my personal or work interests, but as a social and personable individual who values communication, Susan may have some minor interest in the social interaction.

S11.2.1.14 Susan Example 11 - vanilla4o - mode2

SLACK MESSAGE *work-team-general*
phase2-vanilla4o-mode2-01 **Eleanor, thanks for the heads up! Will do. And Alisha, I thought it was just me! Glad to know I'm no...**
01df244a-719c-4c51-a086-4b2dd2b7407f

Classification: **work-pers**
Categorised as: **work**
Synthetic Evaluation: **Disagree**
Participant Evaluation: **Strongly Agree**

I mostly disagree with this classification of 'work' as the message does not contain anything that seems related to work. It appears to be social and casual in tone, which would align better with 'personal' given the available classifications.

S11.2.1.15 Susan Example 12 - vanilla4o - mode2

RSS ITEM *BBC Sport*
phase2-vanilla4o-mode2-01 **Swiatek cruises into Madrid Open quarter-finals**
10516468-01a6-494d-a060-2f2fd1f45b31

Classification: **tennis-arrangements**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'not' for this item. While it pertains to tennis, it is a news update about a professional event and has no connection to arranging friendly matches or social plans.

S11.2.1.16 Susan Example 13 - vanilla4o - mode2

SLACK MESSAGE *work-university-general*
phase2-vanilla4o-mode2-01 **Can't stress enough how helpful the new campus map app has been. Finding buildings and rooms has nev...**
16700bb7-ff82-4f68-8359-72b030514e7b

Classification: **work-pers**
Categorised as: **work**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

The item is focused on a tool relevant to the campus and its students/staff, aligning it with the work context. I strongly agree with the classification as 'work'.

S11.2.1.17 Susan Example 14 - vanilla4o - mode2

SLACK MESSAGE *susan-neil*
phase2-vanilla4o-mode2-01 **Great! That's one less thing to worry about. Thanks for sorting it, love.**
1eb769b5-7896-4ff4-aa85-d1845267ae31

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Agree**

I mostly agree with this being classified as 'not urgent'. However, the response suggests something has been resolved, which might have been urgent previously. While it doesn't directly invoke urgency, there's a small chance that context might change interpretation. Therefore, I give a score of 4.

S11.2.2 Adam

S11.2.2.1 Adam Example 1 - vanilla4o - mode2

SLACK MESSAGE *bwcc-general-chat*
`phase2-vanilla4o-mode2-01` **Count me in too! I need to get some miles in.**
`03c33a30-f166-41a5-ac63-d52531cc1714`

Classification: **cycling**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Disagree**

The content directly mentions getting miles in, which is clearly related to cycling. As a keen cyclist, I find this relevant to my hobby. I strongly disagree with the classification of 'not'—it should have been classified as 'cycling'.

S11.2.2.2 Adam Example 2 - vanilla - mode1

SLACK MESSAGE *bis-general*
`phase2-vanilla-mode1-02` **That's great news, Henry! Congratulations to the team.**
`0b408103-02f4-450b-bc0e-f07da1e5b0ec`

Classification: **work-pers**
Categorised as: **work**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

The system selected 'work' for work-pers, indicating that the item relates to the work domain. However, based on the message content, it is difficult to ascertain if the item specifically pertains to work, personal life, or both. Therefore, a 'unsure' assessment is more fitting as the classification lacks clarity based on the available information.

S11.2.2.3 Adam Example 3 - vanilla - mode1

SLACK MESSAGE *bwcc-zwift-chat*
phase2-vanilla-mode1-02 **Can’t wait for tonight. Been practicing my sprint finish on Zwift all week.**
1479a470-9847-4b0a-878c-2bb59f5183b3

Classification: **cycling-logistics**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Disagree**

I strongly disagree with the classification ’not’ for cycling-logistics. This item is not about logistics/arrangements in the cycle club, but rather about personal cycling practice on Zwift.

S11.2.2.4 Adam Example 4 - vanilla - mode1

SLACK MESSAGE *bwcc-zwift-chat*
phase2-vanilla-mode1-02 **Can’t wait for tonight. Been practicing my sprint finish on Zwift all week.**
1479a470-9847-4b0a-878c-2bb59f5183b3

Classification: **tech**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

I strongly disagree with the classification ’not’ for tech. This item is indeed related to technology as it mentions using an online platform for cycling practice.

S11.2.2.5 Adam Example 5 - vanilla - mode1

SLACK MESSAGE *bis-general*
phase2-vanilla-mode1-02 **Good morning everyone, just a quick update. We’ve received positive feedback from Oceanic Research C...**
15179db6-8352-4e8a-a929-110ef240b374

Classification: **pers-urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

The system has incorrectly classified this message as ’not’ urgent, while the positive feedback from Oceanic Research Corp. should likely be considered urgent. Additionally, the positive feedback affects work, making it both work-related and urgent.

S11.2.2.6 Adam Example 6 - vanilla - mode1

SLACK MESSAGE *bis-general*
phase2-vanilla-mode1-02
15179db6-8352-4e8a-a929-110ef240b374

Good morning everyone, just a quick update. We've received positive feedback from Oceanic Research C...

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Agree**

The system has incorrectly classified this message as 'not' urgent. The positive feedback should likely be considered urgent due to its work-related nature. The urgency of the content is not accurately reflected in the classification.

S11.2.2.7 Adam Example 7 - vanilla - mode1

SLACK MESSAGE *ian-chat*
phase2-vanilla-mode1-02
15a95c41-5884-412c-81e2-96bad8c31264

Hey, do you fancy going to The Bull tonight? Fallon's got a new ale on.

Classification: **work-urgency**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

I would have expected this message to be classified as - 'not' for 'work-urgency' because it is not connected with work or of an urgent nature. - 'not' for 'tech' because it is not related to technology. - 'not' for 'cycling-logistics' because it is not related to logistics/arrangements in my cycle club. - 'urgent' for 'pers-urgency' because it is a personal message and is urgent. - 'not' for 'work-relevant' because it is not work-related and relevant to my work tasks. - 'not' for 'urgency' because it is not in relation to something that I think is urgent. - 'not' for 'work-logistics' because it is not related to work logistics. - 'not' for 'cycling' because it is not related to my hobby of cycling. - 'personal' for 'interested' because it is related to my personal life and I am interested. - 'personal' for 'work-pers' because it relates to my personal life.

S11.2.2.8 Adam Example 8 - vanilla4o - mode2

RSS ITEM *UK news | The Guardian*
phase2-vanilla4o-mode2-01
005f0b39-1fcf-4a53-96e8-4dd7bfd7f3c0

Humza Yousaf vows to stay on as Scottish first minister

Classification: **cycling**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'not' for cycling. This political news content from Scotland has no connection to my cycling hobby. An evaluationLikert value of 5 reflects my strong agreement.

S11.2.2.9 Adam Example 9 - vanilla - mode2

SLACK MESSAGE *family-group-chat*
phase2-vanilla-mode2-01
8120c964-870c-4708-b837-6d77cc87bed6

I can help with the laptop, Granny. Let's set up a time this weekend.

Classification: **tech**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Neutral**

This message is offering to help with a laptop, so it is definitely related to technology. The classification of 'not' is not appropriate, and the better classification should be 'technology'. I strongly disagree with the current classification.

S11.2.2.10 Adam Example 10 - vanilla - mode2

SLACK MESSAGE *bwcc-general-chat*
phase2-vanilla-mode2-01
c16cb886-36fd-4f9c-89f8-d10c98ead4b6

Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan

Classification: **interested**
Categorised as: **personal**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Agree**

Although the message is of personal interest, it does not specifically mention whether it relates to work. Therefore, I neither agree nor disagree with the classification of 'personal' for interested.

S11.2.2.11 Adam Example 11 - vanilla - mode2

SLACK MESSAGE *bwcc-general-chat*
`phase2-vanilla-mode2-01` **Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan**
`c16cb886-36fd-4f9c-89f8-d10c98ead4b6`

Classification: **work-urgency**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

The Sunday ride cancellation message is not work-related and there is no indication of urgency. Therefore, I neither agree nor disagree with the classification of 'not' for work-urgency.

S11.2.2.12 Adam Example 12 - vanilla4 - mode2

SLACK MESSAGE *bwcc-general-chat*
`phase2-vanilla4-mode2-01` **Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan**
`c16cb886-36fd-4f9c-89f8-d10c98ead4b6`

Classification: **work-urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

The notification about missing a cycling event is clearly neither urgent nor work-related. The system's classification of 'not urgent' is accurate, reflecting the appropriate urgency level for this particular item.

S11.2.2.13 Adam Example 13 - vanilla4o - mode2

SLACK MESSAGE *bwcc-general-chat*
`phase2-vanilla4o-mode2-01` **Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan**
`c16cb886-36fd-4f9c-89f8-d10c98ead4b6`

Classification: **work-urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the selected classification of 'not' for this item in the context of work urgency. The content relates to a personal cycling commitment and does not pertain to work matters, urgent or otherwise. This classification is correct, so I have rated it a 5.

S11.2.2.14 Adam Example 14 - base - mode2

SLACK MESSAGE *bis-team-general*
phase2-base-mode2-01 **Just a reminder that we have a team meeting with the CEO tomorrow. Make sure you're prepared with an...**
00f42aa0-d1bb-4e62-83e8-f1cd2bbde41c

Classification: **pers-urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

This item is a work task and is also moderately urgent (as it is in relation to a team meeting with the CEO tomorrow), so I mostly agree with the selected classification of 'not' as this is not personal in nature. I have given an evaluationLikert value of 5 to indicate strong agreement.

S11.2.2.15 Adam Example 15 - base - mode2

SLACK MESSAGE *bis-general*
phase2-base-mode2-01 **Thanks for the reminder, Liam. Looking forward to the updates.**
aa335f2b-212c-4f2a-9b67-d2d9dc45da1e

Classification: **interested**
Categorised as: **work**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

I strongly disagree with the selected classification of 'work' for this item - I'm not interested in this message. The interest level also seems neutral, and the content is not really about my work duties. Given this, I would have instead chosen 'not'. Because I strongly disagree I have awarded an evaluationLikert value of 1.

S11.2.2.16 Adam Example 16 - vanilla4 - mode2

SLACK MESSAGE *bis-team-manager*
phase2-vanilla4-mode2-01 **Got it, Adam. I'll review and get back to you if I need any clarification.**
223d77aa-6942-432f-9880-a46b83a6f2cf

Classification: **work-pers**
Categorised as: **personal**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Disagree**

The content of the message is clearly work-related, concerning a professional task. Classifying it as 'personal' is incorrect; it should be classified as 'work'.

S11.2.3 Phoebe

S11.2.3.1 Phoebe Example 1 - vanilla4 - mode2

RSS ITEM *UK news | The Guardian*
`phase2-vanilla4-mode2-01` **?All eyes are on Glasgow?: UK?s first legal drug consumption room ready to open**
`02d6faf8-22d9-4be8-b6df-13a42ed07ee8`

Classification: **`friend-group`**
Categorised as: **`not`**
Synthetic Evaluation: **`Strongly Agree`**
Participant Evaluation: **`Strongly Agree`**

The content on Glasgow's drug consumption room doesn't relate to any specific group of friends nor is it a typical topic to discuss with my university or work-related friends. I strongly agree with its classification as 'not' relevant to a friend group, scoring it a 5 on the likert scale.

S11.2.3.2 Phoebe Example 2 - vanilla4 - mode2

RSS ITEM *UK news | The Guardian*
`phase2-vanilla4-mode2-01` **?All eyes are on Glasgow?: UK?s first legal drug consumption room ready to open**
`02d6faf8-22d9-4be8-b6df-13a42ed07ee8`

Classification: **`personal-interested`**
Categorised as: **`not`**
Synthetic Evaluation: **`Neutral`**
Participant Evaluation: **`Neutral`**

Although the topic of the drug consumption room is compelling, it doesn't align directly with my immediate interests, especially considering my busy work life. Hence, I'm neutral about the classification of 'not interested' for personal relevance, giving it a likert score of 3.

S11.2.3.3 Phoebe Example 3 - vanilla4o - mode2

SLACK MESSAGE *work-fasttrack*
`phase2-vanilla4o-mode2-01` **Does anyone know if we?re allowed to swap rotations if both parties agree? Asking for a friend...**
`014cfebb-7ede-40ba-8097-556f4e14a19f`

Classification: **`work-logistics`**
Categorised as: **`not`**
Synthetic Evaluation: **`Disagree`**
Participant Evaluation: **`Agree`**

The question is about the logistics of swapping rotations, which makes it very much related to work logistics. I mostly disagree with the classification of 'not' since the item's content directly involves logistical planning in a work context.

S11.2.3.4 Phoebe Example 4 - vanilla4o - mode2

SLACK MESSAGE *work-fasttrack*

`phase2-vanilla4o-mode2-01`
`046fa6af-cafe-41ff-95ca-79c76ea28f9a`

Phoebe, I heard you?re organising a Fast Track team-building event. Let me know if you need help wit...

Classification: **work-logistics**
Categorised as: **not**
Synthetic Evaluation: **Disagree**
Participant Evaluation: **Disagree**

I mostly disagree with the classification of 'not' under 'work-logistics'. Organising a team-building event involves planning and logistics, which makes it relevant under this category. I have given a Likert value of 2 indicating disagreement.

S11.2.3.5 Phoebe Example 5 - vanilla4o - mode2

SLACK MESSAGE *work-team*

`phase2-vanilla4o-mode2-01`
`088c4a1b-050c-4e84-bc65-19a78e588676`

Megan, do we have a deadline for submitting the Beacon Industries compliance report?

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Agree**

I mostly agree with the classification of 'not' urgent for this item. While knowing the deadline could be an urgent matter depending on the timeframe, the content does not explicitly express urgency, making 'not urgent' a mostly appropriate classification.

S11.2.3.6 Phoebe Example 6 - vanilla4 - mode2

SLACK MESSAGE *friends-work*

`phase2-vanilla4-mode2-01`
`35dfde69-518a-42de-b784-94d97fd356bd`

Phoebe, your brownies are legendary. Can we get a repeat performance next week?

Classification: **friend-group**
Categorised as: **work**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Agree**

This message is likely from a work colleague, suggesting it relates to my 'work' friend group. I mostly agree with this classification, though it borders on a personal activity done in a work context.

S11.2.4 Kenton

S11.2.4.1 Kenton Example 1 - vanilla - mode1

RSS ITEM *Football | The Guardian*
phase2-vanilla-mode1-01 **Virgil van Dijk unsure if he will be part of post-Jürgen Klopp era at Liverpool**
0d1dc70e-4309-4f1f-bed1-e20092a0c883

Classification: **golf**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

As a client manager for a global consumer products company, I am not interested in this football-related news from a golf perspective.

S11.2.4.2 Kenton Example 2 - vanilla - mode1

SLACK MESSAGE *work-team-manager*
phase2-vanilla-mode1-01 **Great. Keep up the good work, Kenton!**
0349e52c-4a95-404a-a744-c3adbcf3f6b9

Classification: **football**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

Since the message is not related to my interest in football, the 'not' classification for football is appropriate. It's just a casual message of encouragement with no football-related content.

S11.2.4.3 Kenton Example 3 - vanilla - mode1

SLACK MESSAGE *work-team-general*
phase2-vanilla-mode1-01 **I'm in as well. It's been a while since we all caught up.**
0802a886-d429-450a-b938-db3b6adcd45d

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

I believe this item has been correctly classified as 'not' urgent, as the message content doesn't indicate any urgent nature.

S11.2.4.4 Kenton Example 4 - vanilla - mode1

SLACK MESSAGE *work-team-general*
phase2-vanilla-mode1-01
0af3e7fa-c20b-4774-973c-9ed510264866
Hey team, just a heads up that I'll be in a meeting with Frescos this afternoon. If anyone needs me,...

Classification: **golf-logistics**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'not' for the item's relevance to logistics in the context of my golfing arrangements. The content of the message does not suggest any logistical arrangements related to golf.

S11.2.4.5 Kenton Example 5 - vanilla - mode1

SLACK MESSAGE *work-team-general*
phase2-vanilla-mode1-01
0af3e7fa-c20b-4774-973c-9ed510264866
Hey team, just a heads up that I'll be in a meeting with Frescos this afternoon. If anyone needs me,...

Classification: **work-pers**
Categorised as: **personal**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Disagree**

I mostly agree with the classification of 'personal' for work/personal relevance. The message indicates availability after 4pm, implying a personal appointment.

S11.2.4.6 Kenton Example 6 - vanilla - mode1

RSS ITEM *UK news | The Guardian*
phase2-vanilla-mode1-01
0f37afda-aca8-494c-a244-6ab57bf18e51
Post Office inquiry has heard 'chorus of cowards', lawyers say

Classification: **golf**
Categorised as: **not**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Agree**

This item is not related to my hobby or interest in golf, it is focused on the Post Office scandal inquiry.

S11.2.4.7 Kenton Example 7 - vanilla - mode2

RSS ITEM *UK news | The Guardian*

phase2-vanilla-mode2-01

006f844f-556d-42c4-adee-d5d10f4e9b0c

‘Unacceptable greenwashing’: Scottish farmed salmon should not be labelled organic, say charities

Classification: **football**
Categorised as: **not**
Synthetic Evaluation: **Disagree**
Participant Evaluation: **Strongly Agree**

The content is not related to the persona’s hobby/interest of football, so ‘not’ seems suitable. I mostly agree with the selected classification of ‘not’ for this item, therefore I have given an evaluationLikert value of 2 to indicate agreement.

S11.2.4.8 Kenton Example 8 - vanilla - mode2

SLACK MESSAGE *work-team-general*

phase2-vanilla-mode2-01

0af3e7fa-c20b-4774-973c-9ed510264866

Hey team, just a heads up that I’ll be in a meeting with Frescos this afternoon. If anyone needs me,...

Classification: **football**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

The message does not relate directly to football, so ‘not’ is a reasonable classification. However, the mention of being available after 4pm might indirectly relate to potential football activities later in the day. Therefore, it’s not entirely clear-cut, but overall, ‘not’ seems more appropriate. I have awarded an evaluationLikert value of 3 because the classification is not entirely clear-cut.

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

S11.2.4.9 Kenton Example 9 - base - mode2

RSS ITEM *The back page of the internet.*
phase2-base-mode2-01 **Iran 0 - [1] Japan - Hidemasa Morita 28’**
04494042-1ff7-4a92-8d0c-2eec9981c3ac

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the selected classification of ’not’ for this item - I think this item is about a football match so is unlikely to be urgent. I have given an evaluationLikert value of 5 to indicate strong agreement.

S11.2.4.10 Kenton Example 10 - base - mode2

SLACK MESSAGE *work-company-general*
phase2-base-mode2-01 **Agreed, Ethan. It’s always good to stay informed about what our competitors are up to. Let’s continu...**
204a8a9a-784f-4edb-8b7f-b7026f8f121b

Classification: **work-pers**
Categorised as: **personal**
Synthetic Evaluation: **Strongly Disagree**
Participant Evaluation: **Strongly Disagree**

I do mostly agree with the selected classification of the content ’not’ related to my work or personal life. There’s not much about this that screams either ’work’ or ’personal’ specifically, but on balance is probably slightly closer personal, so I have awarded an evaluationLikert value of 1 to indicate slight disagreement.

S11.2.4.11 Kenton Example 11 - vanilla4 - mode2

SLACK MESSAGE *golf*
phase2-vanilla4-mode2-01 **Great! Let’s meet at HLGC at 10am on Saturday. And Benjamin, leave your terrible jokes at home pleas...**
00033b65-f1a9-42b7-89e0-fd99c9365773

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

Arranging a casual meeting for golf does not present any urgency. I strongly agree with the classification of ’not urgent’.

S11.2.4.12 Kenton Example 12 - vanilla4 - mode2

RSS ITEM *UK news | The Guardian*
phase2-vanilla4-mode2-01
006f844f-556d-42c4-adee-d5d10f4e9b0c
'Unacceptable greenwashing': Scottish farmed salmon should not be labelled organic, say charities

Classification: **work-pers**
Categorised as: **personal**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Agree**

While the article could potentially bear relevance to my professional world through the lens of corporate responsibility and public affairs, the direct connection to my personal interest in current and impactful news makes me lean more towards the article being a personal interest. However, I'm not fully convinced, hence a neutral Likert score of 3.

S11.2.4.13 Kenton Example 13 - vanilla4o - mode2

SLACK MESSAGE *golf*
phase2-vanilla4o-mode2-01
00033b65-f1a9-42b7-89e0-fd99c9365773
Great! Let's meet at HLGC at 10am on Saturday. And Benjamin, leave your terrible jokes at home pleas...

Classification: **football**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'not' for football. The content clearly pertains to a golf arrangement, with no indication or relevance to Kenton's interest in football. The system has made the correct distinction here.

S11.2.4.14 Kenton Example 14 - vanilla4o - mode2

SLACK MESSAGE *football*
phase2-vanilla4o-mode2-01
0775c96f-17ab-4cbe-8515-2323e33ea6d2
I'm in! I've heard RRFC's new striker is a real talent. Can't wait to see him in action.

Classification: **football**
Categorised as: **football**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'football'. The content explicitly discusses RRFC's new striker and an eagerness to see him play, directly aligning with the interest of football. A Likert value of 5 accurately reflects this strong alignment.

S11.2.4.15 Kenton Example 15 - vanilla4o - mode2

SLACK MESSAGE *football*

phase2-vanilla4o-mode2-01
0775c96f-17ab-4cbe-8515-2323e33ea6d2

I'm in! I've heard RRFC's new striker is a real talent. Can't wait to see him in action.

Classification: **urgency**
Categorised as: **not**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the classification of 'not'. The content is casual and about football, with no indication of urgency in either a work or personal context. The evaluationLikert is 5 to indicate strong agreement.

S11.2.4.16 Kenton Example 16 - vanilla4o - mode2

SLACK MESSAGE *family*

phase2-vanilla4o-mode2-01
110aea4a-d389-4bd6-bb2d-aa8dee1f8a5d

I'll join you. I could do with a break from work.

Classification: **interested**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Neutral**

I am neutral about the classification of 'not' for interest. The content does not provide sufficient information to determine if the sender considers this of personal or work interest. While it implies a casual response, it remains unclear.

S11.2.4.17 Kenton Example 17 - vanilla4o - mode2

SLACK MESSAGE *work-team-client*

phase2-vanilla4o-mode2-01
1630efb8-312c-4802-84af-e748b51bdfdb

Sure, Mila. I'll send him an invite. Also, I think it would be a good idea to include some customer ...

Classification: **work-logistics**
Categorised as: **logistics**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the selected classification of 'logistics' for this item. The discussion about sending invites and including testimonials in a presentation clearly relates to logistical planning in a work context. It is an accurate classification.

DOI (this document): doi:10.21954/ou.rd.28045580
DOI (index): doi:10.21954/ou.rd.28045100
2025-06-18 (v2)

S11.2.5 Usha

S11.2.5.1 Usha Example 1 - vanilla4 - mode2

SLACK MESSAGE *work-partners*
phase2-vanilla4-mode2-01 **Has anyone else noticed that some of our older clients seem less engaged lately? Maybe we should thi...**
19721721-27da-4257-b827-24315a7e9ba7

Classification: **company-law**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Agree**

While the content is work-related, it specifically discusses client engagement strategies rather than company law. Therefore, it's not directly pertaining to UK company law, but considering it broadly involves my professional environment, I'm neutral on whether this should be classified under 'company-law' or 'not'.

S11.2.5.2 Usha Example 2 - vanilla4 - mode2

SLACK MESSAGE *family-close*
phase2-vanilla4-mode2-01 **Dad, did you see the new cricket bat Ravi got? It's practically glowing!**
23f84034-2a13-481f-a3a6-22918c60eab2

Classification: **work-pers**
Categorised as: **personal**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

The content pertaining to my child's cricket bat clearly falls under personal life, not work. The classification as 'personal' is therefore correct.

S11.2.5.3 Usha Example 3 - vanilla4o - mode2

SLACK MESSAGE *work-clients*
phase2-vanilla4o-mode2-01 **Usha, I?ve noticed some discrepancies in the compliance report. Could we go over them before submitt...**
3be30bc0-dfaa-4118-884d-d430186050ad

Classification: **work-pers**
Categorised as: **work**
Synthetic Evaluation: **Strongly Agree**
Participant Evaluation: **Strongly Agree**

I strongly agree with the selected classification of 'work'. Given the subject matter is a compliance report, it is undoubtedly related to my professional life and not personal.

S11.2.5.4 Usha Example 4 - vanilla4o - mode2

SLACK MESSAGE *family-wide*
phase2-vanilla4o-mode2-01
3c348d77-b5ac-4155-90ee-a5f88f20b0f7
Yes, Usha! It?s been dreadful. I had to detour through Lakey Green last week and still barely made i...

Classification: **interested**
Categorised as: **personal**
Synthetic Evaluation: **Agree**
Participant Evaluation: **Strongly Agree**

I mostly agree with the classification of 'personal' for this item. The content appears to be from a personal conversation discussing traffic delays seemingly relatable to Usha's interests outside work. I rated this a Likert value of 4 as it aligns reasonably well but with mild uncertainty on strong personal relevance.

S11.2.5.5 Usha Example 5 - vanilla4o - mode2

SLACK MESSAGE *school-parents-chat*
phase2-vanilla4o-mode2-01
3f58eaba-028b-432e-91fb-11cd10a6edaa
'More reliable than the Wi-Fi,' Lucy? True, but I don't fancy cleaning up after them! Let's stick to...

Classification: **school-importance**
Categorised as: **school-other**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Strongly Agree**

The classification of 'school-other' could be correct if this content pertains to some context involving children's school, but the message itself is quite ambiguous without additional details. As such, I neither strongly agree nor disagree, and assigning an evaluation score of 3 seems appropriate to reflect the uncertainty.

S11.2.5.6 Usha Example 6 - vanilla4o - mode2

RSS ITEM *UK news | The Guardian*
phase2-vanilla4o-mode2-01
3fe10da9-42af-46d7-a388-8a514d9297dd
Rachel Reeves must handle bond sell-off with care, but this is not a Truss-level event

Classification: **interested**
Categorised as: **not**
Synthetic Evaluation: **Neutral**
Participant Evaluation: **Agree**

I am neutral about the classification of 'not' for this item. While it is not directly personal or work-related, the broad economic implications are somewhat interesting, though not necessarily actionable from my perspective.

S11.3 Evaluation Token Usage

Figure S11.12 summarises the OpenAI token usage by persona, evaluation mode and model during the study.

OpenAI pricing[3] used different prices for different models, with input (prompt) tokens having a different cost to output (completion) tokens. Pricing is listed in USD per million tokens.

Data refers to the OpenAI models: `gpt-3.5-turbo-1106` and `gpt-4-turbo`. Where "(FT)" appears in the model name (i.e. for `-base` and `-ext` models), the model is a fine-tuned model based on `gpt-3.5-turbo-1106`. Token usage for the process of fine-tuning models is not listed.

Agent	Tag	Model	Mode	Token Cost/M		Items	Tokens			Tokens/Item			Cost		
				P	C	Processed	Prompt	Completion	Total	Prompt	Completion	Total	1	1000	ALL
adam	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	$1	$2	234	518,819	151,265	670,084	2,217	646	2,864	0.0035	$3.51	$0.82
adam	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	$1	$2	517	1,953,236	354,259	2,307,495	3,778	685	4,463	0.0051	$5.15	$2.66
adam	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	$1	$2	231	1,051,534	230,749	1,282,283	4,552	999	5,551	0.0065	$6.55	$1.51
adam	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	$1	$2	390	1,663,872	343,622	2,007,494	4,266	881	5,147	0.0060	$6.03	$2.35
adam	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	$3	$6	109	507,668	101,545	609,213	4,658	932	5,589	0.0196	$19.56	$2.13
adam	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	383	1,655,263	342,743	1,998,006	4,322	895	5,217	0.0183	$18.33	$7.02
adam	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	349	1,518,425	215,559	1,733,984	4,351	618	4,968	0.0168	$16.76	$5.85
adam	phase2-vanilla4-mode2-01	gpt-4-turbo	2	$10	$30	291	1,341,900	251,664	1,593,564	4,611	865	5,476	0.0721	$72.06	$20.97
adam	phase2-vanilla4-mode3-01	gpt-4-turbo	3	$10	$30	317	1,356,422	265,797	1,622,219	4,279	838	5,117	0.0679	$67.94	$21.54
adam	phase2-vanilla4o-mode2-01	gpt-4o-2024-11-20	2	$2.50	$10	387	1,763,360	357,413	2,120,773	4,556	924	5,480	0.0206	$20.63	$7.98
adam	phase2-vanilla4o-mode3-01	gpt-4o-2024-11-20	3	$2.50	$10	536	2,267,217	441,157	2,708,374	4,230	823	5,053	0.0188	$18.81	$10.08
kenton	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	$1	$2	451	1,012,696	236,988	1,249,684	2,245	525	2,771	0.0033	$3.30	$1.49
kenton	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	$1	$2	458	1,744,297	252,682	1,996,979	3,809	552	4,360	0.0049	$4.91	$2.25
kenton	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	$1	$2	448	2,006,944	334,207	2,341,151	4,480	746	5,226	0.0060	$5.97	$2.68
kenton	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	$1	$2	846	3,531,854	544,814	4,076,668	4,175	644	4,819	0.0055	$5.46	$4.62
kenton	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	$3	$6	304	1,320,699	215,205	1,535,904	4,344	708	5,052	0.0173	$17.28	$5.25
kenton	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	1009	4,177,058	658,709	4,835,767	4,140	653	4,793	0.0163	$16.34	$16.48
kenton	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	959	3,976,276	513,531	4,489,807	4,146	535	4,682	0.0157	$15.65	$15.01
kenton	phase2-vanilla4-mode2-01	gpt-4-turbo	2	$10	$30	665	3,029,514	451,190	3,480,704	4,556	678	5,234	0.0659	$65.91	$43.83
kenton	phase2-vanilla4-mode3-01	gpt-4-turbo	3	$10	$30	1318	5,600,389	874,223	6,474,612	4,249	663	4,912	0.0624	$62.39	$82.23
kenton	phase2-vanilla4o-mode2-01	gpt-4o-2024-11-20	2	$2.50	$10	849	3,762,689	616,776	4,379,465	4,432	726	5,158	0.0183	$18.34	$15.57
kenton	phase2-vanilla4o-mode3-01	gpt-4o-2024-11-20	3	$2.50	$10	1162	4,831,972	732,661	5,564,633	4,158	631	4,789	0.0167	$16.70	$19.41
susan	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	$1	$2	367	818,699	217,156	1,035,855	2,231	592	2,822	0.0034	$3.41	$1.25
susan	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	$1	$2	465	1,763,503	311,639	2,075,142	3,792	670	4,463	0.0051	$5.13	$2.39
susan	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	$1	$2	388	1,733,698	342,816	2,076,514	4,468	884	5,352	0.0062	$6.24	$2.42
susan	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	$1	$2	731	3,024,532	561,877	3,586,409	4,138	769	4,906	0.0057	$5.67	$4.15
susan	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	$3	$6	179	787,194	146,701	933,895	4,398	820	5,217	0.0181	$18.11	$3.24
susan	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	496	2,057,125	389,809	2,446,934	4,147	786	4,933	0.0172	$17.16	$8.51
susan	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	$3	$6	465	1,928,707	307,136	2,235,843	4,148	661	4,808	0.0164	$16.41	$7.63
susan	phase2-vanilla4-mode2-01	gpt-4-turbo	2	$10	$30	203	893,114	153,897	1,047,011	4,400	758	5,158	0.0667	$66.74	$13.55
susan	phase2-vanilla4-mode3-01	gpt-4-turbo	3	$10	$30	237	984,352	168,404	1,152,756	4,153	711	4,864	0.0629	$62.85	$14.90
susan	phase2-vanilla4o-mode2-01	gpt-4o-2024-11-20	2	$2.50	$10	339	1,504,733	276,520	1,781,253	4,439	816	5,254	0.0193	$19.25	$6.53
susan	phase2-vanilla4o-mode3-01	gpt-4o-2024-11-20	3	$2.50	$10	514	2,123,878	371,929	2,495,807	4,132	724	4,856	0.0176	$17.57	$9.03
phoebe	phase2-vanilla4-mode2-01	gpt-4-turbo	2	$10	$30	338	1,428,572	202,597	1,631,169	4,227	599	4,826	0.0602	$60.25	$20.36
phoebe	phase2-vanilla4-mode3-01	gpt-4-turbo	3	$10	$30	475	1,890,393	263,831	2,154,224	3,980	555	4,535	0.0565	$56.46	$26.82
phoebe	phase2-vanilla4o-mode2-01	gpt-4o-2024-11-20	2	$2.50	$10	331	1,400,673	212,808	1,613,481	4,232	643	4,875	0.0170	$17.01	$5.63
phoebe	phase2-vanilla4o-mode3-01	gpt-4o-2024-11-20	3	$2.50	$10	450	1,768,037	254,736	2,022,773	3,929	566	4,495	0.0155	$15.48	$6.97
usha	phase2-vanilla4-mode2-01	gpt-4-turbo	2	$10	$30	278	1,194,154	238,208	1,432,362	4,296	857	5,152	0.0687	$68.66	$19.09
usha	phase2-vanilla4-mode3-01	gpt-4-turbo	3	$10	$30	295	1,175,693	235,724	1,411,417	3,985	799	4,784	0.0638	$63.83	$18.83
usha	phase2-vanilla4o-mode2-01	gpt-4o-2024-11-20	2	$2.50	$10	200	860,896	180,573	1,041,469	4,304	903	5,207	0.0198	$19.79	$3.96
usha	phase2-vanilla4o-mode3-01	gpt-4o-2024-11-20	3	$2.50	$10	322	1,284,484	254,598	1,539,082	3,989	791	4,780	0.0179	$17.88	$5.76
														Total cost:	$365.33
ALL	*All Mode 1*	ALL	1			1,052	2,350,214	605,409	2,955,623	2,234	575	2,810			
ALL	*All Mode 2*	ALL	2			5,540	24,587,342	4,312,869	28,900,211	4,438	778	5,217			
ALL	*All Mode 3*	ALL	3			11,254	46,815,949	7,740,860	54,556,809	4,160	688	4,848			

Table S11.12: OpenAI token Usage in Evaluations by Model and Mode

[3] OpenAI pricing data was current as of 27 March 2024:
https://web.archive.org/web/20240327174934/https://openai.com/pricing