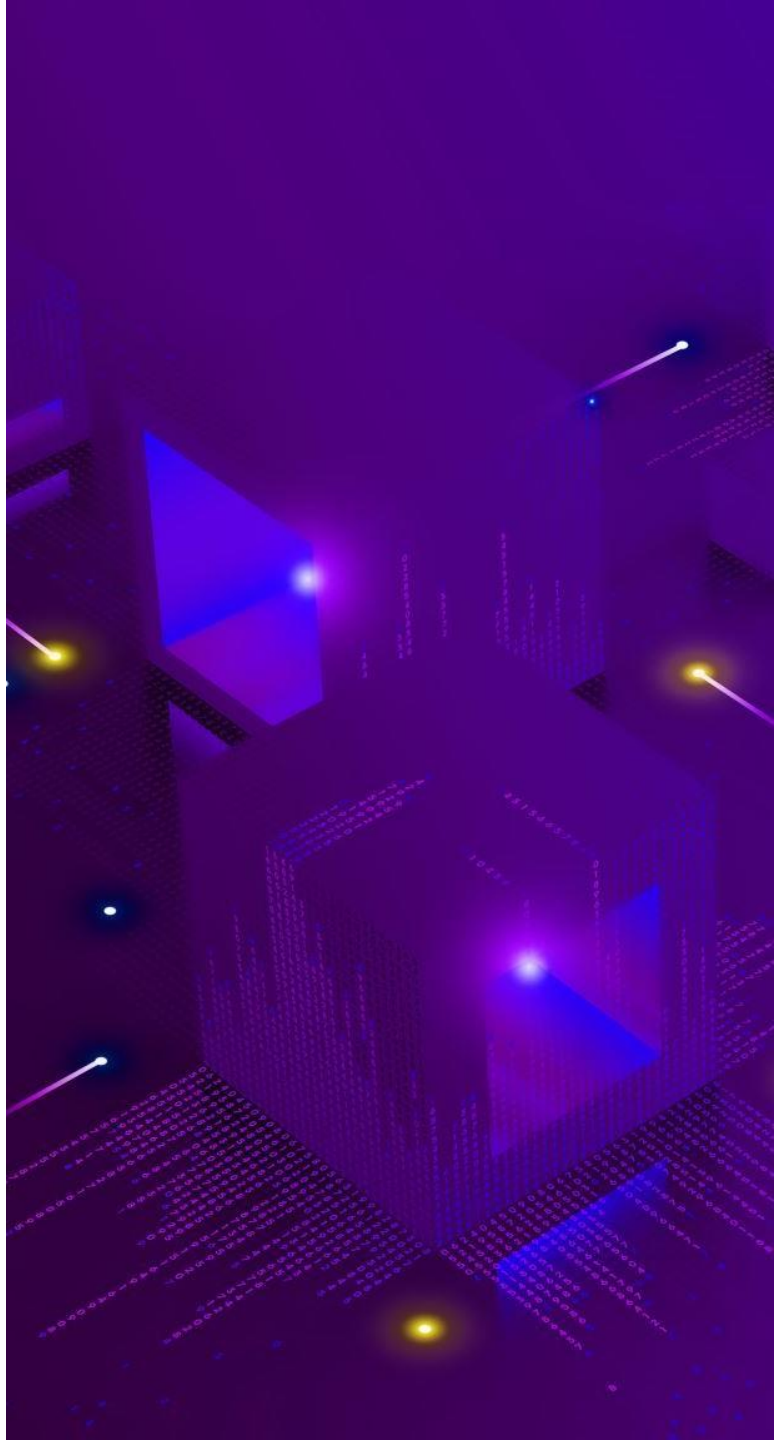
HUMAN EVALUATION AND AUTOMATIC METRICS, CORRELATION ANALYSIS

Presentazione risultati



Sommario

- Framework di analisi
- Analisi sul test set
- Analisi CodeGen
- Analisi CodeT5
- Analisi CodeGPT

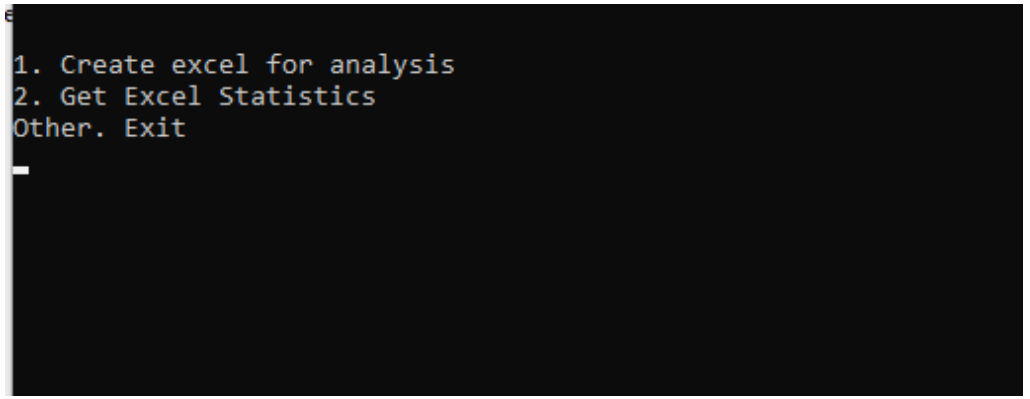
Framework di analisi

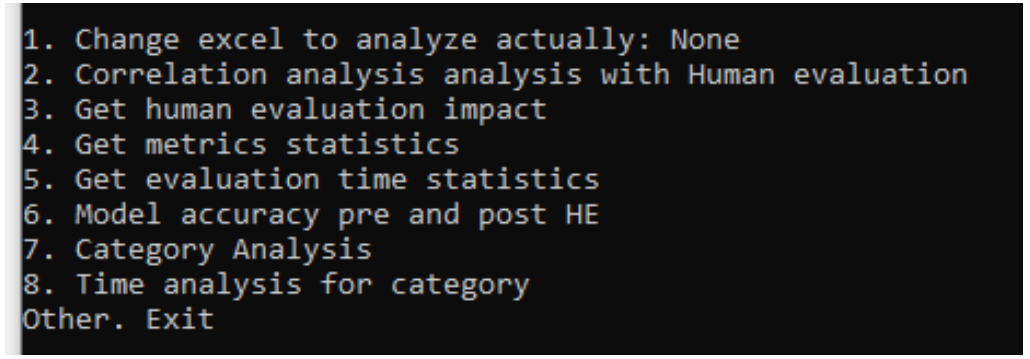
Allo scopo di effettuare analisi statistiche sugli output dei modelli è stato strutturato e documentato un framework di analisi in modo da promuovere la ripetibilità.

Caratteristiche:

- Scritto in python
- Tool a linea di comando
- Interazione semplice
- Documentato con Doxygen

Framework di analisi





Le operazioni possibili al momento sono due

- Creare un excel per effettuare HE e avere le metriche single line
- Ottenere delle statistiche sulle analisi svolte

Tra le statistiche ottenibili ci sono:

- Analisi di correlazione con le metriche con Kendall e Spermann
- Impatto delle HE
- Statistiche riguardo le metriche
- Statistiche sul tempo di valutazione
- Analisi per categoria (sia temporale che di accuracy)

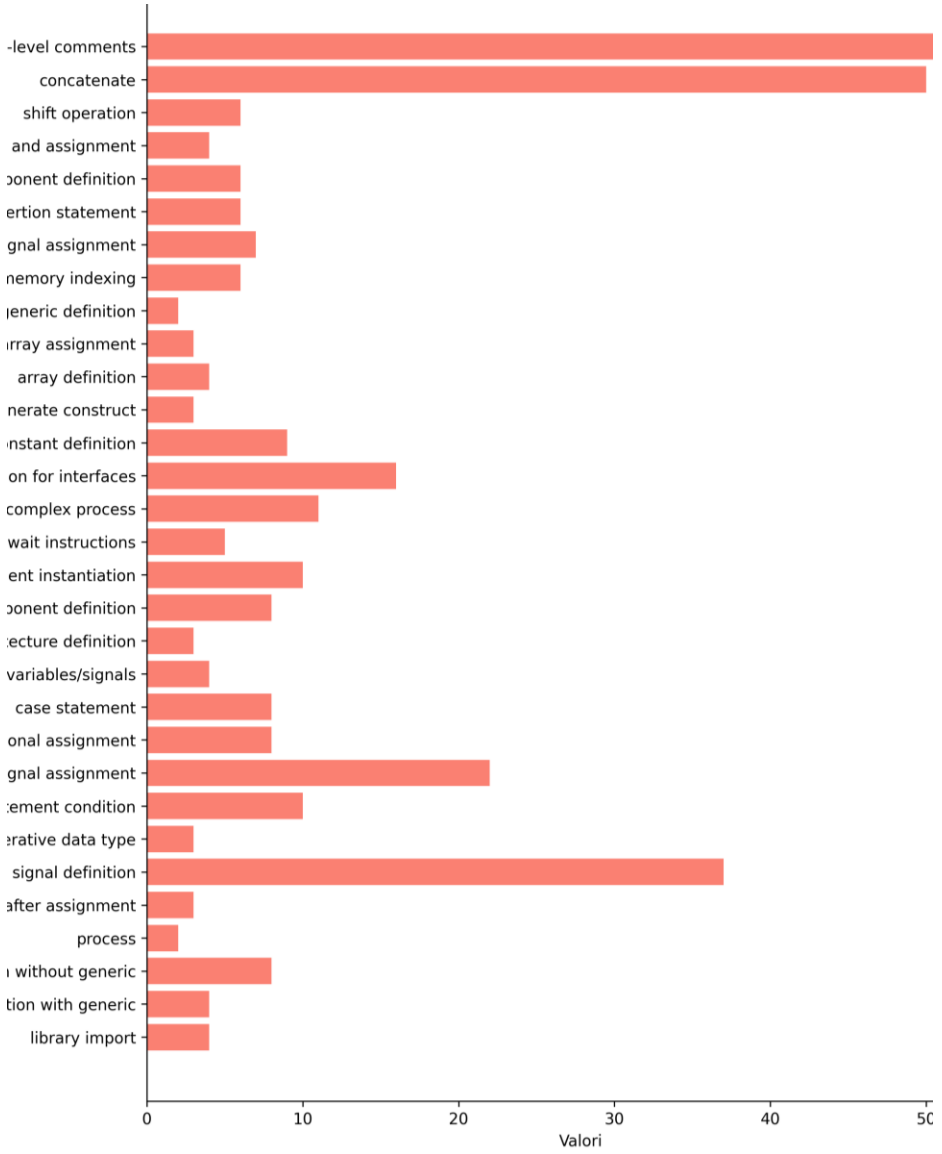
Framework di analisi

Il framework è stato documentato usando doxygen per renderlo facilemente usabile ed estendibile qualora si volessero integrare ulteriori analisi

ANALISI SU TEST SET

Di seguito si riporta la distribuzione delle categorie nel testset.

Possiamo notare che il test set presenta molte entry per le categorie **level comments, concatenate** e **signal definition**



ANALISI CODEGEN

Si riportano di seguito le analisi svolte sul modello Code Gen, in particolare si riportano le analisi sul tempo di valutazione, l'impatto della human evaluation ed eventuali problemi presenti nel dataset

Tempo medio per entry: 83406 ms (1:23:41 minuti:secondi:millisecondi)

Tempo massimo per 5 entry: 632041 ms (10:32:41 minuti:secondi:millisecondi)

Righe relative al tempo massimo: 236,237,238,239,240

Tempo totale di valutazione: 4:4:39:397 (ore:minuti:secondi:millisecondi)

Human evaluation impact: ones_before: 150 ones_after: 219 human evaluation impact: 69

Accuracy pre_HE(ones_before/n_entry**)**: 0.46 **post HE(**ones_after/n_entry**):** 0.67

Problems: Aggiunta random di blocchi di ROM non richiesti

ANALISI CODEGEN (CORRELAZIONE)

Si riporta di seguito l'analisi di correlazione con le metriche previste

	Spermann	P-value(Speramnn)	Kendall	P-value(Kendall)
EM	0.645	<0.05	0.645	<0.05
ED	0.668	<0.05	0.583	<0.05
METEOR	0.590	<0.05	0.485	<0.05
LCS	0.672	<0.05	0.587	<0.05
CRYSTAL BLEU	0.557	<0.05	0.493	<0.05
SACRE BLEU	0.568	<0.05	0.504	<0.05
ROUGE	0.669	<0.05	0.586	<0.05

ANALISI CODEGEN (CATEGORIA)

Si riporta di seguito l'analisi dell'accuracy(correct_entry/number_of_entry) per categoria.

In particolare riportiamo solo le categorie più significative al fine delle analisi

entity definition without generic	**0.875**
process	0
enumerative data type	0.33
if statement condition	0.5
case statement	0.25
architecture definition	0.66
component definition	0.375
component instantiation	0.1
for generate construct	0.33
complex component definition	1
element array assignment	0.33
memory indexing	0.33

ANALISI CODE T5 220

Si riportano di seguito le analisi svolte sul modello Code T5 220, in particolare si riportano le analisi sul tempo di valutazione, l'impatto della human evaluation ed eventuali problemi presenti nel dataset

Tempo medio per entry: 44373 ms (0:44:37 minuti:secondi:millisecondi)

Tempo massimo per 5 entry: 414042 ms (6:54:42 minuti:secondi:millisecondi)

Righe relative al tempo massimo: 271,272,273,274,275

Tempo totale di valutazione: 2:13:7:60 (ore:minuti:secondi:millisecondi)

Human evaluation impact: ones_before: 146 ones_after: 264 human evaluation impact: 118

Accuracy pre_HE(ones_before/n_entry**)**: 0.47 **post HE(**ones_after/n_entry**):** 0.81

Problems: Nessuno

ANALISI CODE T5 220 (CORRELAZIONE)

Si riporta di seguito l'analisi di correlazione con le metriche previste

	Spermann	P-value(Speramnn)	Kendall	P-value(Kendall)
EM	0.436	<0.05	0.436	<0.05
ED	0.630	<0.05	0.550	<0.05
METEOR	0.630	<0.05	0.520	<0.05
LCS	0.631	<0.05	0.550	<0.05
CRYSTAL BLEU	0.616	<0.05	0.548	<0.05
SACRE BLEU	0.499	<0.05	0.448	<0.05
ROUGE	0.570	<0.05	0.498	<0.05

ANALISI CODE T5 220 (CATEGORIA)

Si riporta di seguito l'analisi dell'accuracy(correct_entry/number_of_entry) per categoria.

In particolare riportiamo solo le categorie più significative al fine delle analisi

entity definition without generic	**0.875**
process	0.5
enumerative data type	1
if statement condition	0.9
case statement	0.875
architecture definition	1
component definition	1
component instantiation	0.9
for generate construct	0.66
complex component definition	0.5
element array assignment	1
memory indexing	0.83

ANALISI CODE GPT

Si riportano di seguito le analisi svolte sul modello Code GPT, in particolare si riportano le analisi sul tempo di valutazione, l'impatto della human evaluation ed eventuali problemi presenti nel dataset

Tempo medio per entry: 29069.204 ms (0:0:29:69 ore:minuti:secondi:millisecondi)

Tempo massimo per 5 entry: 372010 ms (0:6:12:10 ore:minuti:secondi:millisecondi)

Righe relative al tempo massimo: 239,240,241,242,243

Human evaluation impact: ones_before: 40 ones_after: 122 human evaluation impact: 82

Accuracy pre_HE(ones_before/n_entry**)**: 0.122 **post HE(**ones_after/n_entry**):**0.374

Problems: Codice in più,ROM random nelle predizioni,EMPTY PRED

ANALISI CODE GPT (CORRELAZIONE)

Si riporta di seguito l'analisi di correlazione con le metriche previste

	Spermann	P-value(Speramnn)	Kendall	P-value(Kendall)
EM	0.483	<0.05	0.483	<0.05
ED	0.565	<0.05	0.465	<0.05
METEOR	0.677	<0.05	0.556	<0.05
LCS	0.550	<0.05	0.453	<0.05
CRYSTAL BLEU	0.503	<0.05	0.427	<0.05
SACRE BLEU	0.510	<0.05	0.425	<0.05
ROUGE	0.589	<0.05	0.487	<0.05

ANALISI CODE GPT (CATEGORIA)

Si riporta di seguito l'analisi dell'accuracy(correct_entry/number_of_entry) per categoria.

In particolare riportiamo solo le categorie più significative al fine delle analisi

entity definition without generic	0.75
process	0.5
enumerative data type	0.333
if statement condition	0.3
case statement	0.125
architecture definition	0.333
component definition	0.125
component instantiation	0.3
for generate construct	0
complex component definition	0.5
element array assignment	0.333
memory indexing	0.333

CORRELAZIONE GLOBALE

Si riporta di seguito la media delle correlazioni sui 3 modelli analizzati

	Spermann	Kendall
EM	0.521	0.521
ED	0.621	0.533
METEOR	0.632	0.520
LCS	0.618	0.530
CRYSTAL BLEU	0.559	0.489
SACRE BLEU	0.526	0.459
ROUGE	0.609	0.524

Grazie dell'attenzione