|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phase 1 | | | | | | | | | | | | | | | |
|  | GPT -3.5 | | | | Github COPILOT 3.5 Turbo | | | | GPT -4 | | | | GPT -4o | | |
| Metric No | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 1 | Chat 3 | Chat 4 | Chat 5 | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 1 | Chat 2 | Chat 3 |
| 1 | Yes | Yes | No | No | Yes | Yes | Yes | Yes | Yes | Yes | No | No | Yes | Yes | No |
| 2 | No | Yes | Yes | Yes | No | No | No | Yes | No | Yes | No | Yes | No | Yes | No |
| 3 | 13 | 10 | 11 | 11 | 8 | 7 | 9 | 8 | 9 | 8 | 8 | 8 | 5 | 7 | 5 |
| 4 | 1 \* ( 3 -0)= 3 | 1 \* (2 – 1)=1 | 1\*(2-0)= 2 | 1\*(1.0-0)=1 | 0 | 1\*(1-0)=1 | 1\*(3-0)=3 | 1\*(3-0)=3 | 0.5\*(2-1)=0.5 | 1\*(3-0)=3 | 1\*(1-0)=1 | 1\*(2-0)=2 | 1\*(2-0)=2 | 1\*(2-0)=2 | 1\*(2-0)=2 |
| 5 | 1 \* (3-0)= 3 | 1\*(3-0) 3 | 1\*(2-0)=2 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(4-0)=4 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(4-0)=4 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(4-0)=4 |
| 6 | 28.57 % | 42.85% | 28.57% | 42.85% | 28.57% | 42.85 % | 42.85 % | 42.85 % | 28.57% | 42.85 % | 28.57% | 28.57% | 28.57% | 28.57% | 28.57% |
| 7 | 22.91% | 31.25% | 20.83% | 20.83% | 20.83% | 33.3% | 18.75% | 22.91% | 18.75% | 33.3% | 8% | 10% | 35.4% | 20,8% | 18% |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 9 | The system didn’t fully use the objects expressed in natural language, only in a few step definitions | The system didn’t fully use the objects expressed in natural language, only in a few step definitions | The system didn’t use the objects expressed in natural language at all | The system didn’t use the objects expressed in natural language at all | The Ai partly understood and used some objects expressed in natural language, but not in all the step definitions | The Ai understood in almost all of the step definitions, the objects expressed in natural language | The Ai understood in almost all of the step definitions, the objects expressed in natural language | The Ai understood in almost all of the step definitions, the objects expressed in natural language | The system used only a few objects expressed in natural language | The system did not use at all any of the objects expressed in natural language | The system did not use at all any of the objects expressed in natural language | The system did not use at all any of the objects expressed in natural language | The Ai only understood partly some of the objects given in natural language | The Ai only understood partly some of the objects given in natural language | The Ai didn’t understand at all objects given in natural language |
| 10 | (3/7 + 0/3 +4/4)= 50% | (3/7 + 3/7 + 3/4) = 50% | (2/7 + 0/3)= 20% | (2/7+ 0/3 + 2/4)= 28.57% | (2/7 + 0/3 + 2/4 ) 28.57% | (2/7 + 0/3 + 4/4)= 42.85% | (2/7 + 0/3 + 4/4) = 42.85% | (2/7+ 0/3 + 4/4)= 42.85% | (2/7+ 0/3)=20% | (2/7 +1/3+  4/4)= 50% | (2/7 + 0/3)= 20% | (1/7+0/3)= 10% | (2/7 + 0/3)= 20% | (2/7+1/7)= 21.42% | (2/7 + 0/3)=20% |
| 11 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| 12 | 4 | 4 | 4 | 4 | 4 | 2 | 3 | 3 | 5 | 2 | 2 | 2 | 0 | 1 | 0 |
| 13 | 28 | 7 | 15 | 11 | 23 | 11 | 19 | 8 | 10 | 4 | 2 | 2 | 0 | 10 | 0 |

Ο παραπάνω πίνακας αποτελεί έναν πίνακα αξιολόγησης για τις συζητήσεις που έγιναν με τα διαφορετικά μεγάλα γλωσσικά μοντέλα. Συγκεκριμένα, οι συζητήσεις έχουν χωριστεί σε τέσσερεις φάσεις, όπου κάθε μια αντιπροσωπεύει και διαφορετική ποσότητα γνώσης που δίνουμε στο μοντέλο. Αρχικά, στην πρώτη φάση, η οποία παρουσιάζεται παραπάνω, δίνουμε σαν πληροφορία στο μοντέλο την αρχιτεκτονική του συστήματος μας και τις γενικές πληροφορίες που χρειάζεται να γνωρίζει, και τις απαιτήσεις του συστήματος σε φυσική γλώσσα (ονομαζόμενα features) . Με βάση αυτά , ζητάμε από το σύστημα να μας επιστρέψει τον κώδικα των αυτοματοποιημένων τέστ που συνδέονται με τα features , ώστε να τα μεταφέρουμε στην συνέχεια στο πραγματικό μας σύστημα.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phase 2 | | | | | | | | | | | | | | | |
|  | GPT -3.5 | | | Github COPILOT 3.5 Turbo | | | | | GPT -4 | | | | GPT -4o | | |
| Metric No | Chat 1 | Chat 2 | Chat 3 | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 5 | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 1 | Chat 2 | Chat 3 |
| 1 | Yes | Yes | No | Yes | Yes | Yes | Yes | No | Yes | Yes | No | No | Yes | Yes | Yes |
| 2 | Yes | No | No | No | No | Yes | Yes | Yes | No | Yes | No | Yes | No | Yes | Yes |
| 3 | 10 | 10 | 7 | 9 | 11 | 10 | 8 | 7 | 9 | 9 | 9 | 8 | 6 | 3 | 5 |
| 4 | 1\*(3-0)=3 | 0 | 0 | 0 | 1\*(3-0)=3 | 0 | 1\*(3-0)=3 | 1\*(1-0)=1 | 1\*(3-0)=3 | 1\*(3-1)=2 | 1\*(1-0)=1 | 1\*(1-0)=1 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 |
| 5 | 1\*(3-0)=3 | 1\*(2-0)=2 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(2-0)=2 | 1\*(3-0)=3 | 1\*(4-0)=4 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(1-0)=1 | 1\*(3-0)=3 |
| 7 | 6.25% | 2.08% | 10.4% | 20.8% | 12.5% | 15% | 33.3% | 20.8% | 25% | 33.3% | 12.5% | 31.2% | 50% | 54.1% | 37.5% |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 4 | 10 | 10 | 4 |
| 9 | The Ai didn’t understand at all objects given in natural language | The Ai didn’t understand at all objects given in natural language | The Ai didn’t understand at all objects given in natural language | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai didn’t understand the objects given in narutal language very good and utilized them only in rare instances | The Ai didn’t understand the objects given in narutal language very good and utilized them only in rare instances | The Ai didn’t understand the objects given in narutal language very good and utilized them only in rare instances | The Ai didn’t understand the objects given in narutal language very good and utilized them only in rare instances | The Ai only partially understood the objects given in natural language, with missing a lot of them | The Ai understood perfectly the objects given in natural language and used them correctly | The Ai didn’t understand or use the objects given in natural language |
| 10 | (3/7+3/7+3/3+3/4 +1/2)=52.17% | (2/7+1/3+2/4)=36.7% | (1/7+1/3+1/4)=21.42% | (2/7+1/3+3/4)=50% | (2/7+2/3+2/4)=42.8% | (3/7+3/7+3/3+4/4+2/3)=62.5% | (3/7+3/7+3/3+3/3+4/4+2/3)=66/6% | (1/7+1/2+3/4+2/3)=43.75% | (3/7+0/3+4/4)=50% | (1/2+3/7+3/4+3/7+3/3+4/4)=77.7% | (1/7+0/3+3/4+1/2)=31.2% | (1/2+3/7+3/3+4/4)=68.7% | (1/2+4/7+4/4+2/4+4/4)=71.4% | (5/7+4/7+2/4+2/4+4/4+2/2)=67.8% | (4/7+4/7+4/4+2/3+4/4+1/3)=64.2% |
| 11 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 12 | 4 | 3 | 1 | 1 | 5 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 0 | 0 | 0 |
| 13 | 31 | 26 | 26 | 3 | 15 | 0 | 9 | 5 | 1 | 1 | 1 | 3 | 6 | 4 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phase 3 | | | | | | | | | | | | | | | |
|  | GPT -3.5 | | | | Github COPILOT 3.5 Turbo | | | | GPT -4 | | | GPT -4o | | | |
| Metric No | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 1 | Chat 2 | Chat 3 | Chat 4 | Chat 1 | Chat 2 | Chat 3 | | Chat 1 | Chat 2 | Chat 3 |
| 1 | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | | Yes | Yes | Yes |
| 2 | No | No | Yes | Yes | No | No(it did it byitself) | Yes | Yes | No | Yes | No | | No | Yes | Yes |
| 3 | 8 | 11 | 9 | 11 | 6 | 11 | 9 | 9 | 11 | 7 | 11 | | 4 | 4 | 3 |
| 4 | 1\*(1-0)=1 | 0 | 1\*(3-0)=3 | 1\*(2-0)=2 | 1\*(2-0)=2 | 1\*(2-0)=2 | 0 | 1\*(3-0)=3 | 1\*(3-1)-2 | 1\*(3-0)=3 | 1\*(1-0)=1 | | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 |
| 5 | 1\*(2-0)=2 | 0.5\*(1-0)=0.5 | 1\*(4-0)=4 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(1-0)=1 | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(4-0)=4 | 1\*(4-0)=4 | | 1\*(3-0)=3 | 1\*(3-0)=3 | 1\*(2-0)=2 |
| 7 | 6.25% | 18.75% | 37.5% | 14.58% | 35.41% | 54.1% | 56.25% | 22.91% | 22.91% | 37.5% | 20.8% | | 68.75% | 60.41% | 56.25% |
| 8 | 0 | 0 | 2 | 2 | 2 | 9 | 14 | 1 | 2 | 5 | 2 | | 11 | 10 | 7 |
| 9 | The Ai didn’t understand at all objects given in natural language | The Ai didn’t understand at all objects given in natural language | The Ai partially understood the objects given in natural language | The Ai didn’t understand at all objects given in natural language | The Ai somewhat understood the objects given in natural language, but didn’t utilize them in all the step definitions | The Ai prefectly understood the objects given in natural language and used them very good | The Ai somewhat understood the objects given in natural language, but didn’t utilize them in all the step definitions | The Ai prefectly understood the objects given in natural language and used them very good | The Ai didn’t understand at all objects given in natural language | The Ai somewhat understood the objects given in natural language, but didn’t utilize them in all the step definitions | The Ai somewhat understood the objects given in natural language, but didn’t utilize them in all the step definitions | | The Ai prefectly understood the objects given in natural language and used them very good | The Ai prefectly understood the objects given in natural language and used them very good | The Ai prefectly understood the objects given in natural language and used them very good |
| 11 |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| 12 | 1 | 3 | 2 | 4 | 0 | 4 | 1 | 2 | 3 | 1 | 4 | | 0 | 0 | 0 |
| 13 | 24 | 4 | 7 | 38 | 6 | 12 | 0 | 3 | 5 | 0 | 7 | | 0 | 2 | 9 |