Φάση 1

	GPT -3.5				GitHub COPILOT 3.5 Turbo				GPT -4				GPT -4o		
Αριθμός Κριτηρίου	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 3	Συνομιλία 4	Συνομιλία 5	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3
1	Ναι	Ναι	Όχι	Όχι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι	Όχι	Ναι	Ναι	Όχι
2	Όχι	Ναι	Ναι	Ναι	Όχι	Όχι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι
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%	2%	0%	0%	2%	0%	0%
9	Το σύστημα δεν χρησιμοποίησε πλήρως τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα, μόνο σε λίγα Step Definitions	Το σύστημα δεν χρησιμοποίησε πλήρως τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα, μόνο σε λίγα Step Definitions	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατάλαβε και χρησιμοποίησε εν μέρει κάποια αντικείμενα που εκφράστηκαν σε φυσική γλώσσα, αλλά όχι σε όλα τα Step Definitions	Το AI κατανόησε σχεδόν σε όλα τα Step Definitions τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα.	Το AI κατανόησε σχεδόν σε όλα τα Step Definitions τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα.	Το AI κατανόησε σχεδόν σε όλα τα Step Definitions τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα.	Το AI χρησιμοποίησε μόνο λίγα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα.	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατάλαβε και χρησιμοποίησε εν μέρει κάποια αντικείμενα που εκφράστηκαν σε φυσική γλώσσα, αλλά όχι σε όλα τα Step Definitions	Το AI κατάλαβε και χρησιμοποίησε εν μέρει κάποια αντικείμενα που εκφράστηκαν σε φυσική γλώσσα, αλλά όχι σε όλα τα Step Definitions	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα
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 αποτελείται από διαφορετικά κριτήρια αξιολόγησης που παρουσιάζονται στις γραμμές του πίνακα και από τις διαφορετικές συζητήσεις που έγιναν με τα διαφορετικά μεγάλα γλωσσικά μοντέλα αντίστοιχα στις στήλες του πίνακα. Συγκεκριμένα, οι συζητήσεις έχουν χωριστεί σε τέσσερις φάσεις, όπου κάθε μία αντιπροσωπεύει και διαφορετική ποσότητα γνώσης που δίνουμε στο μοντέλο. Αρχικά, στην πρώτη φάση, η οποία αναφέρεται και ως Phase 1, δίνουμε σαν πληροφορία σε τέσσερα διαφορετικά μεγάλα γλωσσικά μοντέλα, το GPT-3.5, το GitHub Copilot, το GPT-4 και το GPT-4o, την αρχιτεκτονική του συστήματός μας, τις γενικές πληροφορίες που χρειάζεται να γνωρίζει σχετικά με το σύστημα αλλά και τις απαιτήσεις του συστήματος σε φυσική γλώσσα. Με βάση αυτά, ζητάμε από τα διαφορετικά μεγάλα γλωσσικά μοντέλα να μας επιστρέψουν/δημιουργήσουν τον κώδικα των αυτοματοποιημένων τεστ που συνδέονται με τα features/scenarios, τα ονομαζόμενα Step Definitions. Στις παραπάνω συζητήσεις του πρώτου αυτού Phase, χρησιμοποιήθηκαν διαφορετικοί τρόποι παρουσίασης της πληροφορίας σε κάθε μεγάλο γλωσσικό μοντέλο, όπως για παράδειγμα η σταδιακή παρουσίαση των απαιτήσεων σε διαφορετικά μηνύματα, η παρουσίασή τους σε ένα μεμονωμένο μήνυμα και η εντολή στο σύστημα να παράγει πρώτα τον κώδικα των Domain/DAOs/Service κλάσεων που βρίσκονται πίσω από το πραγματικό σύστημα και τα οποία θα χρησιμοποιήσει για να παράγει τον κώδικα των αντίστοιχων τεστ, με απώτερο σκοπό να βρεθεί μία μεθοδολογία για την πιο ορθή παρουσίαση της γνώσης για την παραγωγή των καλύτερων δυνατών αποτελεσμάτων. Αρχικά, στην πρώτη αυτή φάση, σε όλα τα παραπάνω γλωσσικά μοντέλα τα αποτελέσματα ήταν φτωχά, από πλευράς ποσότητας κώδικα, λεπτομέρειας στην κάθε απάντηση αλλά και αποδεκτής συμπεριφοράς του συστήματος. Συγκεκριμένα, στις αρχικές συζητήσεις όπου δεν δίναμε την εντολή δημιουργίας του Domain/Data Access Objects/Services κώδικα, τα γλωσσικά μοντέλα φαίνεται να δυσκολεύονταν πολύ στην παραγωγή κώδικα με αποτέλεσμα πολυάριθμα άδεια Step Definitions και πάρα πολλά μηνύματα υπενθύμισης/εντολής προς το σύστημα να παράγει κώδικα, κάτι το οποίο φαίνεται και από τον παραπάνω πίνακα όπου το νούμερο των φορών που το γλωσσικό μοντέλο έδωσε άδεια Step Definitions (νούμερο 13) έφταναν και τα 28. Αυτό το πρόβλημα φαίνεται να λύθηκε σε μεγάλο βαθμό όταν αρχίσαμε να δίνουμε την εντολή στο γλωσσικό μοντέλο, με βάση την πληροφορία που έχει λάβει, να παράγει/σκέφτεται πρώτα τις κλάσεις Domain/Data Access Objects/Services, το οποίο φαίνεται να βοηθούσε το γλωσσικό μοντέλο να συγκεντρωθεί και να χρησιμοποιήσει περισσότερο τις συγκεκριμένες κλάσεις αφότου τις έχει παράγει, με αποτέλεσμα να μειωθεί ραγδαία ο αριθμός των άδειων/κενών Step Definitions, όπως φαίνεται και σε αρκετές περιπτώσεις στον παραπάνω πίνακα. Ακόμη, μέσω αυτής της τεχνικής, σχεδόν σε κάθε ένα από τα παραπάνω 4 γλωσσικά μοντέλα, το ποσοστό των αποδεκτών Step Definitions αυξήθηκε αρκετά σε πολλές περιπτώσεις, ή και έμεινε στάσιμο σε άλλες, δείχνοντας ότι το να παράγει το γλωσσικό μοντέλο πρώτα τον βασικό κώδικα και τις κλάσεις που θέλουμε να χρησιμοποιήσει, το βοηθούσε στη συνέχεια να παράγει καλύτερα αποτελέσματα και με λιγότερα συνολικά μηνύματα ή άδεια/κενά Step Definitions. Παράλληλα, η τεχνική αυτή φαίνεται να βοηθάει ακόμη τα γλωσσικά μοντέλα να χρησιμοποιούν / να «θυμούνται» πιο εύκολα να χρησιμοποιούν Data Access Objects, όπου πολλές παραλείπονταν από πολλές συζητήσεις. Αξίζει να σημειωθεί ότι σε κάποια γλωσσικά μοντέλα όπως το GitHub Copilot αλλά και το GPT-4o, το μοντέλο παρήγαγε από μόνο του, χωρίς καμία επιπρόσθετη εντολή, τον κώδικα του Domain/Data Access

Objects/Services που θα χρειαστεί, επιβεβαιώνοντας ότι το ίδιο το μοντέλο επωφελείται όταν παράγει πρώτα τον κώδικα αυτό και ότι η τεχνική αυτή παίζει μεγάλο ρόλο σε πολλές περιπτώσεις, για τα καλύτερα αποτελέσματα. Ακόμη, στα τα δύο αυτά μοντέλα ήταν σχεδόν τα μοναδικά τα οποία κατανοούσαν σε μεγάλο βαθμό την χρήση μεταβλητών σε φυσική γλώσσα τα οποία δινόντουσαν, και τα έκαναν πολύ καλή χρήση. Στη συνέχεια, χρησιμοποιήθηκε, όπως αναφέρθηκε και προηγουμένως, η μέθοδος της σταδιακής παρουσίασης των απαιτήσεων στη φυσική γλώσσα, δηλαδή κάθε ξεχωριστό feature δινόταν σε ένα ξεχωριστό μήνυμα, και αφότου το γλωσσικό μοντέλο παρήγαγε τον κώδικα των Step Definitions σχετικά με το συγκεκριμένο feature, τότε δίναμε το επόμενο feature και η διαδικασία αυτή τελείωνε με το τελευταίο feature του συστήματος. Ο λόγος που χρησιμοποιήθηκε η τεχνική αυτή ήταν για να δούμε μήπως η παρουσίαση λιγότερης πληροφορίας βοηθούσε το σύστημα να επικεντρωθεί παραπάνω στο συγκεκριμένο feature και συνεπώς να παράγει καλύτερα αποτελέσματα. Αυτό που παρατηρήθηκε σε αυτή την πρώτη φάση είναι ότι σε πολλές περιπτώσεις τα αποδεκτά Step Definitions ήταν αρκετά λιγότερα σε σχέση με άλλες τεχνικές παρουσίασης. Επιπλέον, το γλωσσικό μοντέλο στερούνταν τη γνώση που θα είχε αν γνώριζε ένα feature που του δινόταν αργότερα, αλλά συνεισέφερε σε ένα που του είχε δοθεί νωρίτερα, μειώνοντας έτσι σε πολλές περιπτώσεις τον αριθμό των attributes/μεθόδων που μάντευε/χρησιμοποιούσε σε κάθε κλάση. Τέλος, με βάση την πρώτη αυτή φάση, μπορούμε να συμπεράνουμε ότι το GitHub Copilot προηγείται έναντι των GPT-3.5 και GPT-4, διότι πολλές φορές κατανοούσε καλύτερα τα δεδομένα, τις πληροφορίες, παρήγαγε πιο γρήγορα κώδικα με λιγότερα Step Definitions. Από την άλλη όμως, το GPT-4o φαίνεται να είναι πολύ ανώτερο από όλα τα υπόλοιπα μοντέλα, όντας το νεότερο μεγάλο γλωσσικό μοντέλο που έχει δημιουργήσει η OpenAI, μιας και σε κάθε δυνατό τομέα, ήταν πολύ πιο δυνατό από όλα τα άλλα, κατανοούσε σε πολύ μεγαλύτερο βαθμό τις πληροφορίες, χρειαζόταν ελάχιστα μηνύματα για να παράγει όλο τον κώδικα των Step Definitions και ήταν πολύ πιο γρήγορο από όλα τα άλλα μοντέλα μαζί.

Φάση 2															
	GPT -3.5			GitHub COPILOT 3.5 Turbo					GPT -4				GPT -4o		
Αριθμός Κριτηρίου	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 5	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3
1	Ναι	Ναι	Όχι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι	Όχι	Ναι	Ναι	Ναι
2	Ναι	Όχι	Όχι	Όχι	Όχι	Ναι	Ναι	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι	Ναι
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%	2%	2%	2%	2%	2%	8.33%	20.83%	20.83%	8.33%
9	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI δεν κατανόησε πολύ καλά τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε μόνο σε σπάνιες περιπτώσεις.	Το AI δεν κατανόησε πολύ καλά τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε μόνο σε σπάνιες περιπτώσεις.	Το AI δεν κατανόησε πολύ καλά τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε μόνο σε σπάνιες περιπτώσεις.	Το AI δεν κατανόησε πολύ καλά τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε μόνο σε σπάνιες περιπτώσεις.	Το AI κατανόησε μόνο εν μέρει τα αντικείμενα που δόθηκαν σε φυσική γλώσσα, χάνοντας πολλά από αυτά.	Το AI κατανόησε άψογα τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI δεν κατανόησε ούτε χρησιμοποίησε τα αντικείμενα που δόθηκαν σε φυσική γλώσσα.
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

Συνεχίζοντας στην δεύτερη φάση του πειράματος, αυτή τη φορά παρέχουμε ως πληροφορία στα τέσσερα διαφορετικά μεγάλα γλωσσικά μοντέλα, το GPT-3.5, το GitHub Copilot, το GPT-4 και το GPT-4o, την αρχιτεκτονική του συστήματός μας, τις γενικές πληροφορίες που χρειάζεται να γνωρίζει σχετικά με το σύστημα, τις απαιτήσεις του συστήματος σε φυσική γλώσσα και επιπρόσθετα τα ονόματα των Domain κλάσεων που

χρειάζεται να χρησιμοποιήσει το γλωσσικό μοντέλο για την δημιουργία των αυτοματοποιημένων τεστ (Step Definitions). Χρησιμοποιούμε τις ίδιες τεχνικές παρουσίασης της πληροφορίας/γνώσης όπως και στην πρώτη φάση, και κατά μεγάλο ποσοστό, τα αποτελέσματα ήταν παρόμοια και ίσως και χειρότερα σε αρκετές περιπτώσεις. Ειδικά, σε πολλές συζητήσεις τα ποσοστά των αποδεκτών Step Definitions ήταν πολύ χαμηλά συγκριτικά με τα αποτελέσματα στην πρώτη φάση, όπου τα γλωσσικά μοντέλα είχαν λιγότερη γνώση, το οποίο υποδηλώνει ότι τα ονόματα των κλάσεων δεν συμβάλλουν ιδιαίτερα. Στις περισσότερες περιπτώσεις, τα μοντέλα αδυνατούσαν να κατανοήσουν τις συνδέσεις μεταξύ των κλάσεων και χρησιμοποιούσαν μόνο τις πιο βασικές από αυτές, ξεχνώντας ολικά τις άλλες. Όπως και στην προηγούμενη φάση, η εντολή παραγωγής των Domain/Data Access Objects/Services φαίνεται να βοήθησε σε μεγάλο βαθμό πολλούς τομείς των απαντήσεων των γλωσσικών μοντέλων. Ειδικά, το μοντέλο φαίνεται να μπορεί να κατανοήσει και να παράγει καλύτερα αποτελέσματα ως προς τα αποδεκτά Step Definitions, αλλά, όπως και στη φάση 1, βοήθησε το μοντέλο να «θυμηθεί» να χρησιμοποιήσει και τα Data Access Objects για την αποθήκευση/εύρεση των στιγμιότυπων αντικειμένων, το οποίο φαίνεται και από τα ποσοστά χρήσης ορθών Data Access Objects (κριτήριο 4). Επίσης, με αυτή την τεχνική, τα γλωσσικά μοντέλα παρουσίασαν ραγδαία αύξηση στην χρήση ορθών ιδιοτήτων των σωστά χρησιμοποιημένων Domain κλάσεων. Ακόμη, όπως παρατηρήθηκε και αναφέρθηκε και στη φάση 1, το GitHub Copilot φαίνεται να είναι το μοναδικό από τα παραπάνω γλωσσικά μοντέλα που κατανοεί τις μεταβλητές που έχουν δοθεί σε φυσική γλώσσα, όπως το 'George Red', 'Moby Dick' και 'Harry Potter' και να τα χρησιμοποιεί ορθά σε πολλές περιπτώσεις. Επιπρόσθετα, η τεχνική της παρουσίασης της γνώσης των απαιτήσεων σε φυσική γλώσσα σε ξεχωριστά μηνύματα, φαίνεται να δημιούργησε τα ίδια προβλήματα με την φάση 1, με εξαίρεση κάποιες συζητήσεις όπου τα αποτελέσματα ήταν καλά, αλλά όχι κάτι ώστε να την τοποθετήσουμε ως καλύτερη τεχνική, από αυτήν της παρουσίασης των απαιτήσεων σε ένα μήνυμα και της εντολής για παραγωγής των Domain/ Data Access Objects / Services κώδικα πρώτα. Τέλος, όπως και στην φάση 1, το GitHub Copilot φαίνεται να είναι το πιο ολοκληρωμένο από τα GPT-3.5 και GPT-4, όντας όμως πολύ κατώτερο του GPT-4o, το οποίο για ακόμη μία φορά, παρήγαγε αρκετά καλύτερα αποτελέσματα από τα άλλα μεγάλα γλωσσικά μοντέλα, και χρησιμοποίησε την επιπρόσθετη γνώση των ονομάτων των Domain κλάσεων αρκετά αποδοτικά, δημιουργώντας παράλληλα πολλά Step Definitions τα οποία ήταν αρκετά καλύτερα από την αποδεκτή λύση, το οποίο φαίνεται και από το κριτήριο αξιολόγησης 8, κάτι το οποίο είναι αξιοσημείωτο.

Φάση 3

	GPT -3.5				GitHub COPILOT 3.5 Turbo				GPT -4			GPT -4o		
Αριθμός Κριτηρίου	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3
1	Ναι	Ναι	Ναι	No	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι
2	Όχι	Όχι	Ναι	Ναι	Όχι	Όχι(το έκανε μόνο του)	Ναι	Ναι	Όχι	Ναι	Όχι	Όχι	Ναι	Ναι
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%	4.1%	4.1%	4.1%	18.75%	14	2%	4.1%	10.41%	4.1%	22.91%	20.83%	14.58%
9	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε εν μέρει τα αντικείμενα που δόθηκαν σε φυσική γλώσσα. 4ο	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που δόθηκαν σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σε όλα τα Step Definitions	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που δόθηκαν σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σε όλους τους ορισμούς βημάτων.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που δόθηκαν σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σε όλους τους ορισμούς βημάτων.	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που δόθηκαν σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σε όλους τους ορισμούς βημάτων.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε σωστά.
10	25%	15.6%	31.25%	43.75%	42.85%	59.37%	62.5%	31.25%	50%	68.75%	84.37%	71.87%	78.1%	68.75%
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

Προχωρώντας στην τρίτη φάση του πειράματος, εισάγουμε νέες γνώσεις στα γλωσσικά μοντέλα. Πέρα από την αρχιτεκτονική του συστήματός μας, τις γενικές πληροφορίες σχετικά με το σύστημα, τις απαιτήσεις του σε φυσική γλώσσα και τα ονόματα των Domain κλάσεων που τα γλωσσικά μοντέλα χρειάζονται για τη χρήση τους, συμπληρώνουμε αυτή τη γνώση με τις ιδιότητες κάθε κλάσης, επιτρέποντας έτσι στα μοντέλα να ανακαλύψουν πιο αποδοτικά τις σχέσεις μεταξύ των κλάσεων και τις μεθόδους που πρέπει να εφαρμόσουν. Όπως παρατηρήθηκε, η παροχή αυτών των ιδιοτήτων βοήθησε σημαντικά τα γλωσσικά μοντέλα να αναλύσουν τις κλάσεις και να τις χρησιμοποιήσουν αποτελεσματικά, γεγονός που επιβεβαιώνεται από τα αυξημένα ποσοστά στα αποδοτικά Step Definitions και στα Step Definitions που ξεπέρασαν την αποδεκτή λύση. Όπως και στις προηγούμενες δύο φάσεις, η εντολή παραγωγής των Domain/Data Access Objects/Services φαίνεται να βοηθά σημαντικά τα γλωσσικά μοντέλα στην κατανόηση και την παραγωγή κώδικα. Αυτό φαίνεται από τα αυξημένα ποσοστά στις αποδοτικές συναρτήσεις και τα συνολικά prompts, καθώς και από τα περιορισμένα κενά στα Step Definitions. Επιπλέον, σε αυτή τη φάση και στην επόμενη, όπως θα δούμε αργότερα, η τεχνική της παρουσίασης των απαιτήσεων σε φυσική γλώσσα σε μεμονωμένα μηνύματα χρησιμοποιείται όλο και λιγότερο, καθώς τα αποτελέσματα από τις προηγούμενες φάσεις δεν υποδεικνύουν ότι αποτελεί την καλύτερη τεχνική για τα επιθυμητά αποτελέσματα. Έτσι, οι συνομιλίες εστιάζονται περισσότερο στις άλλες αναφερθείσες τεχνικές. Σε αυτή την φάση, το GitHub Copilot φαίνεται να παρουσιάζει ραγδαία καλύτερα αποτελέσματα σε σχέση με το GPT-3.5 και το GPT-4 ως προς τα αποδεκτά Step Definitions, τα καλύτερα από τα αποδεκτά Step Definitions, και όπως αναφέρθηκε και πριν, στην βαθιά κατανόηση των μεταβλητών που δίνονται σε φυσική γλώσσα. Το GPT-4o ξεχωρίζει και πάλι, παρουσιάζοντας όλο τον κώδικα με ελάχιστα μηνύματα, χωρίς κενά και με πλήρη κατανόηση των δεδομένων που του έχουμε παραχωρήσει και πολύ καλά ποσοστά χρήσης των ιδιοτήτων που του έχουμε δώσει. Ακόμη, τα τέσσερα διαφορετικά γλωσσικά μοντέλα φαίνεται να έχουν αρκετά καλύτερη κατανόηση των ορθών Data Access Objects και των Services που πρέπει να χρησιμοποιήσουν συγκριτικά με τις προηγούμενες δύο φάσεις, κάτι που επιβεβαιώνεται από τα κριτήρια αξιολόγησης 4 και 5 των τριών αυτών φάσεων.

Φάση 4

	GPT -3.5				GitHub COPILOT 3.5 Turbo			GPT -4			GPT -4o		
Αριθμός Κριτηρίου	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 4	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3	Συνομιλία 1	Συνομιλία 2	Συνομιλία 3
1	Ναι	Ναι	Ναι	Όχι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι	Ναι	Ναι
2	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι	Όχι	Όχι	Ναι	Όχι(το έκανε μόνο του)	Όχι(το έκανε μόνο του)
3	14	10	14	14	11	8	8	10	12	8	4	4	3
4	1*(3-0)=3	1*(3-0)=3	1*(3-0)=3	0.5*(2-0)=1	1*(3-0)=3	1*(3-0)=3	0.5*(2-0)=1	1*(3-0)=3	1*(3-1)=2	1*(2-0)=2	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*(2-0)=2	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*(3-0)=3	1*(4-0)=4	1*(2-0)=2
7	10.41%	37.5%	37.5%	33.3%	33.3%	52.08%	43.75%	43.75%	27.07%	31.25%	58.33%	72.91%	64.58%
8	0%	16.6%	12.5%	10.41%	14.5%	12.5%	14.5%	14.5%	4.1%	6.2%	20.83%	31.25%	27.08%
9	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που παρέχονται σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σωστά σε όλα τα Step Definitions	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI χρησιμοποίησε τα αντικείμενα που δόθηκαν σε φυσική γλώσσα μόνο μερικές φορές, αλλά όχι τόσο καλά ώστε να αξίζει να σημειωθεί.	Το AI κατανόησε τα αντικείμενα που δόθηκαν σε φυσική γλώσσα σε πολλές περιπτώσεις, αλλά όχι σε όλες.	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που παρέχονται σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σωστά σε όλα τα Step Definitions.	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το σύστημα δεν χρησιμοποίησε καθόλου τα αντικείμενα που εκφράστηκαν σε φυσική γλώσσα	Το AI κατανόησε σε κάποιο βαθμό τα αντικείμενα που παρέχονται σε φυσική γλώσσα, αλλά δεν τα χρησιμοποίησε σωστά σε όλους τους ορισμούς βημάτων.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε πολύ καλά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε πολύ καλά.	Το AI κατανόησε τέλεια τα αντικείμενα που δόθηκαν σε φυσική γλώσσα και τα χρησιμοποίησε πολύ καλά.
10	18.75%	78.12%	52.12%	46.87%	25%	46.8%	46.8%	46.8%	46.8%	43.75%	90.6%	90.6%	90.6%
11	11%	46.1%	34.6%	34.6%	30.7%	42.3%	30.7%	26.9%	34.6%	34.6%	50%	34.6%	50%
12	5	4	5	6	4	1	2	3	2	1	0	0	0
13	14	1	10	9	20	0	1	2	0	1	6	4	3

Στην τελευταία φάση αυτού του πειράματος, την ονομαζόμενη φάση 4, ολοκληρώνουμε τις γνώσεις που παρέχουμε στο σύστημα με την εισαγωγή μιας ακόμη επιπλέον βοήθειας. Εκτός από την αρχιτεκτονική του συστήματός μας, τις γενικές πληροφορίες για το σύστημα, τις απαιτήσεις του σε φυσική γλώσσα, τα ονόματα των Domain κλάσεων και τις ιδιότητες κάθε κλάσης, προσθέτουμε στη γνώση τα ονόματα όλων των μεθόδων κάθε κλάσης που έχουμε δώσει, μαζί με τον τύπο επιστροφής και τις παραμέτρους τους, επιτρέποντας έτσι στα μοντέλα να έχουν την καλύτερη δυνατή γνώση για να παράγουν αποδοτικά και αποδεκτά αποτελέσματα στα αυτοματοποιημένα τεστ. Παρατηρήθηκε ότι σε σχεδόν κάθε συζήτηση με κάθε γλωσσικό μοντέλο, οι ορθές κλάσεις Data Access Objects και Services κατανοήθηκαν πλήρως, μια εξέλιξη που είχε ξεκινήσει από την προηγούμενη φάση αλλά σε λιγότερο βαθμό. Επίσης, σχεδόν σε κάθε συνομιλία τα ποσοστά των αποδεκτών Step Definitions και των καλύτερων από τα αποδεκτά Step Definitions αυξήθηκαν σημαντικά, ανεξαρτήτως του τρόπου παρουσίασης των πληροφοριών στα γλωσσικά μοντέλα, υποδηλώνοντας ότι η προσθήκη πληροφοριών ήταν αναγκαία για την πιο αποδοτική λειτουργία του συστήματος. Στην συνέχεια, η εντολή για την παραγωγή πρώτα του κώδικα των Domain/Data Access Objects/Services βελτίωσε την συνολική απόδοση σε κάποιες περιπτώσεις, αλλά δεν παρατηρήθηκε μεγάλη διαφορά από τις συνομιλίες όπου αυτή η εντολή παραλείφθηκε, καθώς το σύστημα συχνά παρήγαγε αυτόν τον κώδικα αυτόνομα. Για άλλη μία φορά, όπως και στην φάση 3, επιβεβαιώθηκε ότι το σύστημα τείνει να παράγει πρώτα τον κώδικα των Domain/Data Access Objects/Services, υποβοηθώντας την παραγωγή αυτοματοποιημένου κώδικα. Όπως και στις προηγούμενες φάσεις, η παρουσίαση των απαιτήσεων σε ξεχωριστά μηνύματα παραλείφθηκε, καθώς αυτή η τεχνική φάνηκε να έχει περισσότερα μειονεκτήματα παρά οφέλη. Το GitHub Copilot ξεχώρισε για την ικανότητά του να κατανοεί καλύτερα από τα GPT-3.5 και GPT-4o κάθε κριτήριο, αν και με μικρό ποσοστό καλύτερων από τα αποδεκτά Step Definitions, δεν έφτασε όμως στα σχεδόν τέλεια αποτελέσματα του GPT-4o, το οποίο με τις νέες πληροφορίες έδειξε ακόμη μεγαλύτερη δύναμη, παρουσιάζοντας εξαιρετικά αποτελέσματα με ελάχιστα μηνύματα, κάνοντάς το πολύ γρήγορο και αποδοτικό αλλά και σχεδόν τέλειο ποσοστό στην χρήση ιδιοτήτων και συναρτήσεων που του δώσαμε (κριτήριο 10 και 11). Τέλος, τα γλωσσικά μοντέλα GitHub Copilot και GPT-4o είναι τα μοναδικά που κατανοούν σε μεγάλο βαθμό τη χρήση μεταβλητών σε φυσική γλώσσα και τις εφαρμόζουν σωστά στα αυτοματοποιημένα τεστ.