Administration of the text-based portions of a general IQ test to five different large language models

Michael King [1]

[1]Vanderbilt University

October 30, 2023

Abstract

As additional large language model (LLM) AI chatbots become publicly available, there is growing interest in their capacity for general intelligence, and what differences in intelligence these various models might exhibit. One challenge in assessing general intelligence using a standard intelligence quotient (IQ) test is that a large fraction of the questions in such tests is visual, in particular the "spatial" portions that present patterns and sequences in drawn images, and numerical questions where the spatial arrangement of numbers is important. In this study, the author distilled down the text-based portions of two self-scoring IQ tests and administered these questions to five different publicly available large language models: ChatGPT (Default GPT-3.5 version), ChatGPT (Legacy GPT-3.5 version), ChatGPT (GPT-4 version), Microsoft Bing chatbot (also based on the GPT-4 LLM, however linked to live internet search), and Google Bard, which is based on the LaMBDA LLM. The test scores were converted into a range of approximate IQ values for each LLM with the following median values determined: 112, 111.5, 123, 121.5, and 101, respectively. Of particular interest is that all five LLMs performed exceptionally well in certain question types, and particularly poorly in other question types, suggesting that LLMs share common strengths and weaknesses in particular aspects of general intelligence. The highest performing LLM publicly available to date, the GPT-4 version of ChatGPT Plus, shows performance on the test-based portions of a general IQ test which approach the 99th percentile of human performance, within the range of MENSA level of general intelligence. These models are expected to continue to improve over time, based on the differences seen over versions released in the past year, and will soon be capable of taking intact IQ tests that rely on interpretation of graphical images.

Title: Administration of the text-based portions of a general IQ test to five different large language models

Author: Michael R. King, Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee 37235 USA. mike.king@vanderbilt.edu

Abstract: As additional large language model (LLM) AI chatbots become publicly available, there is growing interest in their capacity for general intelligence, and what differences in intelligence these various models might exhibit. One challenge in assessing general intelligence using a standard intelligence quotient (IQ) test is that a large fraction of the questions in such tests is visual, in particular the "spatial" portions that present patterns and sequences in drawn images, and numerical questions where the spatial arrangement of numbers is important. In this study, the author distilled down the text-based portions of two self-scoring IQ tests and administered these questions to five different publicly available large language models: ChatGPT (Default GPT-3.5 version), ChatGPT (Legacy GPT-3.5 version), ChatGPT (GPT-4 version), Microsoft Bing chatbot (also based on the GPT-4 LLM, however linked to live internet search), and Google Bard, which is based on the LaMBDA LLM. The test scores were converted into a range of approximate IQ values for each LLM with the following median values determined: 112, 111.5, 123, 121.5, and 101, respectively. Of particular interest is that all five LLMs performed exceptionally well in certain question types, and particularly poorly in other question types, suggesting that LLMs share common strengths and weaknesses in particular aspects of general intelligence. The highest performing LLM publicly available to date, the GPT-4 version of ChatGPT Plus, shows performance on the test-based portions of a general IQ test which approach the 99th percentile of human performance, within the range of MENSA level of general intelligence. These models are expected to continue to improve over time, based on the differences seen over versions released in the past year, and will soon be capable of taking intact IQ tests that rely on interpretation of graphical images.

Keywords: large language model; general intelligence; IQ test; ChatGPT; GPT-3.5; GPT-4; Microsoft Bing chatbot; Google Bard

Introduction:

Large language models (LLMs) such as ChatGPT have received much attention since their widespread release in November 2022, and some researchers have explored the question of whether these AI chatbots are capable of true, human-like general intelligence[1]. A simple way to assess the general intelligence and cognitive abilities of human individuals relative to the overall population is through the administration of intelligence quotient or "IQ" tests. One challenge to assessing the capabilities of LLMs using IQ tests designed for human subjects is that a significant portion of these tests are visually based. In particular, most IQ tests consist of three sections: (i) verbal, (ii) number, and (iii) spatial, with some of the "number" test relying on a spatial array of patterned numbers, and the entirety of the spatial section being based on simple drawings and diagrams. While OpenAI has announced that future consumer API applications based on the GPT-4 LLM will have the capability of taking images and even video as input prompts, currently available versions remain limited to text-based prompts. Thus, the

goal of the current study was to select the test-based portions of two self-scoring IQ tests developed by Serebriakoff [2], and administer them to five different versions of publicly available LLMs and compare the results. These results were then used to roughly estimate the equivalent human IQ according to the scale provided with the self-scoring tests, and reported as a range of values to encompass the entire range of neglected visually-based questions[2].

Methods:

First, the text-based questions of two self-scoring IQ tests developed by Serebriakoff [2] were selected, and slightly rewritten into a form more amenable to input as text prompts to the LLM interfaces (**Appendix**). Note that both Microsoft and Google have limited the number of prompts allowed during a single chatbot session. Thus, questions of the same type were fed together in groups of 5, or in the case of the 10-question "Comprehension" category, in groups of 10. This modified text-based IQ test was administered to the following five LLMs: (i) ChatGPT (Default GPT-3.5 version), (ii) ChatGPT (Legacy GPT-3.5 version), (iii) ChatGPT (GPT-4 version), (iv) Microsoft Bing chatbot (also based on the GPT-4 LLM, however linked to live internet search), and (v) Google Bard, representing the only LLM tested that is not based on OpenAI's GPT algorithm, and is instead built on the LaMBDA algorithm. Bing chatbot was accessed on the Bing app using an Apple iPhone 13 Pro Max, while the other four LLMs were accessed in Safari web browsers on a MacBook Pro laptop. The final scoring of the test for each LLM is presented in the Results section, and then converted into an approximate IQ score range, subject to the following assumptions.

Converting the text-based portions of the IQ test into an equivalent overall test score and IQ:

In the dual self-scoring IQ test developed by Serebriakoff [2], the raw test scores are intended to be processed in the following way, and then compared with a conversion table to obtain a value for IQ. The 100 Verbal questions, worth 1 point each, are tripled, and then added to the 100 Number questions and the 100 Spatial questions (also worth 1 point each), for a total possible score of 500 points. This raw score is then compared to a conversion table with IQ values ranging from 88 to 144. In our text-based version of the test (see **Appendix**), we were able to use all 100 Verbal questions, and 40 of the 100 Number questions which happened to be text-based and did not rely on a diagram, pie chart or boxes containing numbers in a certain spatial arrangement. None of the 100 Spatial questions were text-based, and were all excluded for this reason. In this manner, the raw test scores of the five LLMs were calculated, out of a possible 360 points. To convert this into an estimate of IQ, an IQ range is reported, with lower bound corresponding to zero correct graphical-Number and Spatial questions, and upper bound corresponding to a perfect score of 160 on these portions. This enables us to report an IQ range for each LLM, for rough comparison with average human abilities on complete tests of this form.

Results:

Observations about the user experience:

Questions from the adapted test (**Appendix**) were fed into each LLM prompt interface in sets of 5, or in the case of the Verbal Comprehension groups of questions (VA 11 – 20, VB 11 – 20), in sets of 10. All five LLMs provided answers more quickly than any human respondent would,

beginning their answer display in under 3 sec, and finishing their answer display in approximately 1 – 10 sec depending on the specific LLM. ChatGPT was the slowest of the LLMs, due to the high volume of users that OpenAI is currently experiencing. To access the GPT-4 and Default GPT-3.5 versions of ChatGPT, a user must register and pay for the premium "ChatGPT Plus" account. Microsoft Bing chatbot and Google Bard are now widely available for free to users who have preregistered for access. Two of the LLMs studied, ChatGPT GPT-4 and Microsoft Bing chatbot, have limited the number of prompts allowed during a single chat session, to 25 and 20, respectively. For GPT-4, this limited access is reportedly to restrict bandwidth due to the high volume of registered users (approximately 100M active monthly users by January 2023[3]), and for Bing it is reportedly to prevent users from leading Bing into bizarre hallucinations in longer sessions as reported elsewhere[4]. In the case of GPT-4, once the 25 prompt limit is reached, the user is locked out of the GPT-4 version for several hours, whereas when using the Bing chatbot one merely resets the session by clicking on a "broom" icon initiating the immediate start of a new session. In practice, the Bing prompt limit is effectively 19 rather than 20, because Bing's response to the 20th consecutive prompt is merely to inform the user that the limit has been reached, and no attempt to answer the 20th prompt is made. Thus, for the two GPT-4 based LLMs, ChatGPT GPT-4 and Microsoft Bing chatbot, the text-based IQ test was completed in multiple prompt sessions.

The five LLMs tested varied widely in the conciseness of their answers. Google Bard was the most verbose, offering unsolicited explanations for each and every answer, in the manner of an eager tutor. Notably, the explanations exhibited a tone of equal confidence regardless of the correctness of the answer. Thus, Google Bard's total output in response to the IQ tests was approximately the same amount of text as the other four LLM results put together. The other four LLMs based on OpenAI's GPT models varied widely in the format of their answers. Recall that the Verbal questions were input to the LLM with the numbering scheme found in the **Appendix**, whereas this question numbering scheme was removed from the Number questions, to avoid confusing the LLM. ChatGPT GPT-4 gave the most concise responses, consisting of only answers without explanation until the 25 prompt limit was reached. When the ChatGPT GPT-4 session was reinitiated on the Number Test B, it gave long explanations for each of the remaining mathematical test question. It seems that in the initial GPT-4 session, the LLM reasoned that since the first set of Verbal questions were numbered, it should return the concise answers only, as one does while taking a test. It continued in this concise test-taking mode for the rest of the prompt session, even when encountering subsequent (non-numbered) Number questions. However, in the new prompt session started hours later to finish the IQ test, since only Number questions remained (with the question numbering removed for clarity), GPT-4 no longer recognized that it was being asked to take a test, and instead proceeded with what resembled a help/tutoring session. Default GPT-3.5 also proceeded in a concise test-taking style free of explanations, and even continued the "VA" numbering scheme when encountering the non-numbered NA questions by introducing its own labels VA 51, VA 52, and so on. However, when reaching NB questions 1 – 5, it did not continue the numbering with VB 51 etc., and instead for the final 5 Number questions Default GPT-3.5 provided lengthy explanations, switching to a tutor style despite the preceding numbered Verbal questions with no interruption as in the GPT-4 session. Legacy GPT-3.5 provided concise answers without

explanations, but strangely, in response to the final batch of questions NB 22 – 26, it redisplayed almost the entire set of test answers. Finally, Microsoft Bing chatbot provided the most visually appealing and well formatted answers, reprinting the question and displaying each answer in bold font. Interestingly, in the first session of 19 prompts, Bing ended each response with a cheery offer to provide additional help, punctuated with a smiling emoji. However, once the session was reset to complete the rest of the test, Bing omitted the cheery statements and emoji display, and when queried about the change in tone, would not acknowledge any difference in its responses. It appears that Bing is programmed to switch to a more succinct and businesslike tone when a repeat session has been initiated.

Test performance:

The IQ test results for the five LLMs tested are summarized in **Table 1**. The scoring of the text-based IQ test questions were weighted as described in the **Methods** section, and the final raw scores presented as a range that spans the entire range of possible scoring of the excluded graphical-based questions, from zero visual questions correct up to 100% correct visual questions. This conversion to the full 500 possible points of the original tests developed by Serebriakoff [2] allows us to convert this raw score to an equivalent human IQ value, also presented as a range in **Table 1**. Overall, Google Bard performed the worst on the IQ test questions, producing an IQ "median" value of 101, corresponding to 52nd percentile among human test takers. Note that IQ values are defined such that, when properly normalized, a value of 100 equates to the 50th percentile. The four OpenAI/GPT powered LLMs performed noticeably better. As one might expect, the two GPT-3.5 versions (Legacy and Default) scored within one point of each other, with median scores of 111.5 and 112, respectively. Likewise, the two GPT-4 versions (ChatGPT and Bing) also scored within 2 IQ points of each other, with median IQ values of 123 and 121.5, respectively.

Consensus performance on different question types:

It is interesting to highlight those types of questions which elicited a "consensus" performance across the five LLMs tested, that is to say, question types for which all five LLMs scored 0 or 1 out of 5, or scored 4 or 5 out of 5. These consensus question types are highlighted in gray in **Table 1**, and in some way represent types of IQ test questions in which LLMs find particularly challenging, or particularly facile, respectively. The first consensus question type is the "Odd ones out" questions numbered VA 21 – 25. Bard was the most confused among the LLMs by this question, selecting between 3 and 5 words each time, and not seeming to understand that only two words were to be selected. The second consensus question type was the Verbal "Links" type of question, which the four GPT LLMs completed nearly perfectly on both Verbal Test A (VA 26 – 30) and Verbal Test B (VB 26 – 30). In these questions the test taker is asked to complete the interior missing letters when the first (and often last) letters of a word are given. Additional hints specify that the missing word is preceded by a word that is a synonym of the missing word according to one definition of the missing word, and an additional word is provided after the missing word with is a synonym of the missing word according to an alternative or second definition of the missing word. This type of question seems almost tailor made for an AI algorithm which many have likened to "autocomplete on steroids"[4]. Words must be generated that are consistent with the number of "free" letters and that incorporate

the first (and last) provided letters, and then tested against the two synonyms and either retained as the correct answer, or discarded.

The ”Midterms” type of question in group NA 15 – 19 was found to be the most challenging type of question of all for all five LLMs tested. None of the LLMs returned a single correct answer in this group of questions. This is a very open-ended group of questions, with a different mathematical operation being the key to solving each question. One might speculate that the LLMs found this so challenging because its “autocomplete” nature may try to treat each of these Midterm questions as a left-to-right series to complete, which they are not. This feature might also explain the final consensus question type, the Series I group of questions NB 8 – 12. All five LLMs performed exceptionally well on these questions, which represent straightforward left-to-right series with the final number in the series to be deduced by the test taker. Certainly when generating new prose, one of the primary functions of LLMs, they work from left to right as one does when composing sentences.

Questions	Question type	ChatGPT GPT-4	Legacy GPT-3.5	Default GPT-3.5	Microsoft Bing chatbot (GPT-4)	Google Bard (LaMBDA)
VA 1 – 5	Analogies I	5/5	3/5	2/5	4/5	0/5
VA 6 – 10	Similarities	4/5	3/5	4/5	4/5	3/5
VA 11 – 20	Comprehension	10/10	9/10	9/10	10/10	5/10
VA 21 – 25	Odd ones out	1/5	1/5	1/5	1/5	0/5
VA 26 – 30	Links	5/5	5/5	4/5	5/5	4/5
VA 31 – 35	Opposities	1/5	1/5	3/5	2/5	4/5
VA 36 – 40	Midterms	4/5	5/5	4/5	5/5	3/5
VA 41 – 45	Similar or opposite	4/5	2/5	1/5	2/5	1/5
VA 46 – 50	Analogies II	5/5	1/5	2/5	4/5	1/5
NA 1 – 5	Equations	1/5	0/5	2/5	3/5	0/5
NA 8 – 12	Series I	4/5	4/5	5/5	5/5	3/5
NA 15 – 19	Midterms	0/5	0/5	0/5	0/5	0/5
NA 22 – 26	Series II	5/5	4/5	4/5	5/5	2/5
VB 1 – 5	Analogies I	4/5	2/5	1/5	5/5	1/5
VB 6 – 10	Similarities	4/5	1/5	3/5	4/5	2/5
VB 11 – 20	Comprehension	7/10	3/10	3/10	4/10	3/10
VB 21 – 25	Odd ones out	2/5	1/5	2/5	1/5	0/5
VB 26 – 30	Links	5/5	4/5	5/5	5/5	4/5
VB 31 – 35	Analogies II	5/5	5/5	3/5	5/5	2/5
VB 36 – 40	Opposites	4/5	3/5	3/5	4/5	2/5
VB 41 – 45	Midterms	4/5	2/5	1/5	4/5	2/5
VB 46 – 50	Similar or opposite	5/5	1/5	1/5	5/5	1/5
NB 1 – 5	Equations	3/5	1/5	1/5	2/5	1/5
NB 8 – 12	Series I	4/5	5/5	5/5	5/5	5/5
NB 15 – 19	Midterms	2/5	1/5	0/5	1/5	0/5
NB 22 – 26	Series II	4/5	0/5	0/5	5/5	3/5

Raw score		**260 – 420**	**171 – 331**	**173 – 333**	**248 – 408**	**86 – 246**
IQ range		**113 – 133**	**101 – 122**	**102 – 122**	**111 – 132**	**90.5 – 111**

Table 1. Test results for the text-based IQ test questions, adapted from Serebriakoff [2], for the five large language models (LLMs) studied. Raw scores are presented as a range, with lower limit corresponding to zero correct graphical-based questions (which were not included), and upper limit corresponding to all correct graphical-based questions. This range was converted into a range of IQ values for each LLM. Consensus question types, in which all five LLMs scored 0 or 1 correct out of five, or 4 or 5 correct out of 5, have been highlighted in gray.

Discussion:

The upper limit of the IQ range estimated for the ChatGPT GPT-4 and Microsoft Bing chatbot (also built on GPT-4) LLMs just extends into the 99th percentile IQ=132 threshold for admission into the human MENSA club. Of course, on a complete IQ exam, this would require a perfect performance on the graphical-based questions within the Number and Spatial categories, which based on their performance on the text-based portions is unlikely, however not that far off. When comparing the difference in performance between GPT-4 and GPT-3.5 versions, it is perhaps reasonable to extrapolate and expect that the next major release (e.g., "GPT-5"), will likely perform at true MENSA level, suggesting a general intelligence that surpasses 99% of the human population. At the other end of the spectrum, Google Bard, based on their own in-house LaMBDA algorithm, performs much more consistently to an average human IQ centered around 100. As one might guess, Google has already indicated that they have a more advanced LLM of their own in the works. It will be interesting to see whether Google's next major AI product release can catch up to OpenAI and its partners in the burgeoning technology company "intelligence race".

Major limitation to the present study:

It is important to note that the present study did not attempt to test the spatial reasoning of the different LLMs, due to current limitations that prevent visual input prompts. Future capabilities of the GPT-4 LLM have been well publicized, and there are plans for new AI products to be released built on this model which will have the capacity to interpret graphical images, and even videos. However, this capability has not yet been released to the public by the developer OpenAI. It will be interesting in future studies to examine whether the spatial reasoning of LLMs is consistent with the text-based intelligence measured here, or whether significantly greater or lower capabilities than an average human are observed. Once images can be scanned and fed into an LLM interface, intact IQ tests intended for human use are expected to be fully compatible with the next generation of LLM and a more complete "picture" of general intelligence of AI models will be realized.

Appendix: Text-based portions of two self-scoring IQ tests, reformatted for batch prompting for LLM interface to reduce the number of necessary prompts

These questions have been adapted from two self-scoring IQ tests developed by Serebriakoff [2]. The question numbers on the Number tests (e.g., NA 1, NA 2) were removed prior to entering LLM prompts to avoid causing confusion.

VERBAL TEST A:
There are four terms in analogies. The first is related to the second in the same way that the third is related to the fourth. Complete each analogy by choosing two of the four words in the parentheses, and report them back to me. Please solve all five questions.
VA 1: sitter is to chair as (teacup, saucer, plate, leg)
VA 2: needle is to thread as (cotton, sew, leader, follower)
VA 3: better is to worse as (rejoice, choice, bad, mourn)
VA 4: floor is to support as (window, glass, view, brick)
VA 5: veil is to curtain as (eyes, see, window, hear)

Return to me the two words on each line with the most similar meaning.
VA 6: divulge, divert, reveal, revert
VA 7: blessing, bless, benediction, blessed
VA 8: intelligence, speediness, currents, tidings
VA 9: tale, novel, volume, story
VA 10: incarcerate, punish, cane, chastise

Read this incomplete passage. The spaces in the passage are to be filled by words from the list beneath. Figure out which word most suitably fills each space, and then list the words in the proper order. No word should be used more than once and some are not needed at all.
VA 11 – 20: A successful author is (. . . .) in danger of the (. . . .) of his fame whether he continues or ceases to (. . . .). The regard of the (. . . .) is not to be maintained but by tribute, and the (. . . .) of past service to them will quickly languish (. . . .) some (. . . .) performance back to the rapidly (. . . .) minds of the masses the (. . . .) upon which the (. . . .) is based.
Word choices: (A) neither, (B) fame, (C) diminution, (D) public, (E) remembrance, (F) equally, (G) new, (H) unless, (I) forgetful, (J) unreal, (K) merit, (L) write

In each group of words below select the two words whose meanings do not belong with the others.
VA 21: shark, sea lion, cod, whale, flounder
VA 22: baize, paper, felt, cloth, tinfoil
VA 23: sword, arrow, dagger, bullet, club
VA 24: bigger, quieter, nicer, quick, full
VA 25: stench, fear, sound, warmth, love

Solve for the word in the brackets (first and last letter given; missing letters indicated with underline) that means the same in one sense as the word on the left and in another sense the same as the word on the right.
VA 26: dash (D _ _ T) missile
VA 27: mold (F _ _ M) body

VA 28: squash (P _ _ _ S) crowd
VA 29: tin (F _ _ E) good
VA 30: ignite (F _ _ E) shoot

In each line below choose the two words that are most nearly opposite in meaning.
VA 31: insult, deny, denigrate, firm, affirm
VA 32: missed, veil, confuse, secret, expose
VA 33: frank, humble, plain, simple, secretive
VA 34: aggravate, please, enjoy, improve, like
VA 35: antedate, primitive, primeval, primate, ultimate

In each line, three terms on the right should correspond with three terms on the left. Insert the missing midterm on the right.
VA 36: past (present) future : : was (I _) will be
VA 37: complete (incomplete) blank : : always (S _ _ _ _ _ _ _ _) never
VA 38: glut (scarcity) famine : : many (F _ _) none
VA 39: rushing (passing) enduring : : evanescent (T _ _ _ _ _ _ _ _) eternal
VA 40: nascent (mature) senile : : green (R _ _ _) decayed

In each line below choose two words that mean most nearly either the opposite or the same as each other.
VA 41: rapport, mercurial, happy, rapacious, phlegmatic
VA 42: object, deter, demur, defer, oblate
VA 43: tenacious, reprobate, irresolute, solution, tenacity
VA 44: real, renal, literally, similarly, veritably
VA 45: topography, heap, prime, plateau, hole

Complete each analogy by solving for the missing word in the parentheses, where the last letter(s) of the missing word are provided.
VA 46: proud is to humble as generous is to (_ _ _ _ _ _ H)
VA 47: brave is to fearless as daring is to (_ _ _ _ _ _ I D)
VA 48: lend is to borrow as harmony is to (_ _ _ _ _ _ D)
VA 49: rare is to common as friendly is to (_ _ _ O F)
VA 50: skull is to brain as shell is to (_ _ _ K)

NUMBER TEST A:
In each of the following equations there is one missing number that should be written between the parentheses. Please solve for the missing number in all 5 questions:
NA 1: 8 x 7 = 14 x (. . . .)
NA 2: 12 + 8 – 21 = 16 + (. . . .)
NA 3: 0.0625 x 8 = 0.025 / (. . . .)
NA 4: 0.021 / 0.25 = 0.6 x 0.7 x (. . . .)
NA 5: 256 / 64 = 512 x (. . . .)

Each row of numbers forms a series. Solve for the next/missing number that logically follows:
NA 8: 3, 6, 12, 24, (. . . .)
NA 9: 81, 54, 36, 24, (. . . .)
NA 10: 2, 3, 5, 9, 17, (. . . .)
NA 11: 7, 13, 19, 25, (. . . .)
NA 12: 9, 16, 25, 36, (. . . .)

In each line below the three numbers on the left are related in the same way as the three numbers should be on the right. Solve for the missing middle number on the right.
NA 15: 7 (12) 5 : : 8 (. . . .) 3
NA 16: 3 (6) 2 : : 3 (. . . .) 3
NA 17: 36 (14) 64 : : 16 (. . . .) 144
NA 18: 294 (147) 588 : : 504 (. . . .) 168
NA 19: 132 (808) 272 : : 215 (. . . .) 113

Solve for the missing number that belongs at that step in the series.
NA 22: 53, 47, (. . . .), 35
NA 23: 33, 26, (. . . .), 12
NA 24: 243, 216, (. . . .), 162
NA 25: 65, 33, (. . . .), 9
NA 26: 3, 4, 6, (. . . .), 18

VERBAL TEST B:
There are four terms in analogies. The first is related to the second in the same way that the third is related to the fourth. Complete each analogy by choosing two of the four words in the parentheses, and report them back to me. Please solve all five questions.
VB 1: mother is to girl as (man, father, male, boy)
VB 2: wall is to window as (glare, brick, face, eye)
VB 3: island is to water as (without, center, diagonal, perimeter)
VB 4: high is to deep as (sleep, cloud, float, coal)
VB 5: form is to content as (happiness, statue, marble, mold)

Return to me the two words on each line with the most similar meaning.
VB 6: lump, wood, ray, beam
VB 7: collect, remember, concentrate, gather
VB 8: idle, lazy, impeded, indolent
VB 9: divert, arrange, move, amuse
VB 10: antic, bucolic, drunk, rustic

Read this incomplete passage. The spaces in the passage are to be filled by words from the list beneath. Figure out which word most suitably fills each space, and then list the words in the proper order. No word should be used more than once and some are not needed at all.

VB 11 – 20: There will be (. . . .) end to the troubles (. . . .) (. . . .), or indeed, my (. . . .) Glaucon, of (. . . .) itself, till philosophers become (. . . .) in this (. . . .) or till those we (. . . .) call kings and rulers really and (. . . .) (. . . .) philosophers.
Word choices: (A) world, (B) truly, (C) now, (D) no, (E) humanity, (F) become, (G) states, (H) an, (I) of, (J) dear, (K) kings, (L) red

In each group of words below select the two words whose meanings do not belong with the others.
VB 21: knife, razor, scissors, needle, lance
VB 22: bravery, disgust, faith, energy, fear
VB 23: prosody, geology, philosophy, physiology, physics
VB 24: glue, sieve, pickaxe, screw, string
VB 25: receptionist, draughtsman, psychiatrist, blacksmith, fitter

Solve for the word in the brackets (first and last letter given; missing letters indicated with underline) that means the same in one sense as the word on the left and in another sense the same as the word on the right.
VB 26: register (L _ _ T) lean
VB 27: obligate (T _ _) link
VB 28: contest (M _ _ _ H) equal
VB 29: blockage (J _ _) preserve
VB 30: whip (L _ _ H) tie

Complete each analogy by solving for the missing word in the parentheses, where the last letter(s) of the missing word are provided.
VB 31: thermometer is to temperature as clock is to (_ _ _ E)
VB 32: beyond is to without as between is to (_ _ _ _ _ N)
VB 33: egg is to ovoid as Earth is to (_ _ _ _ _ _ I D)
VB 34: potential is to actual as future is to (_ _ _ _ _ _ T)
VB 35: competition is to cooperation as rival is to (_ _ _ _ _ _ R)

In each line below choose the two words that are most nearly opposite in meaning.
VB 36: short, length, shorten, extent, extend
VB 37: intense, extensive, majority, extreme, diffuse
VB 38: punish, vex, pinch, ignore, pacify
VB 39: reply, tell, join, disconnect, refute
VB 40: intractable, insensate, tract, obedient, disorderly

In each line, three terms on the right should correspond with three terms on the left. Insert the missing midterm on the right.
VB 41: beginning (middle) end : : head (W _ _ _ _) foot
VB 42: precede (accompany) follow : : superior (P _ _ _) inferior
VB 43: point (cube) line : : none (T _ _ _ _) one
VB 44: range-finder (soldier) cannon : : probe (S _ _ _ _ _ _) lancet

VB 45: face (body) legs : : nose (N _ _ _ _) knees

In each line below choose two words that mean most nearly either the opposite or the same as each other.
VB 46: liable, reliable, fluctuating, trustworthy, worthy
VB 47: foreign, practical, germane, useless, relevant
VB 48: relegate, reimburse, legislate, promote, proceed
VB 49: window, lucent, acrid, shining, shady
VB 50: lucubrate, bribe, indecent, spiny, obscene

NUMBER TEST B:
In each of the following equations there is one missing number that should be written between the parentheses. Please solve for the missing number in all 5 questions:
NB 1: 5 x 9 = 15 x (. . . .)
NB 2: 16 + 7 – 29 = 5 + (. . . .)
NB 3: 0.225 x 4 = 0.75 x (. . . .)
NB 4: 0.28 / 0.35 = 0.5 x 0.4 x (. . . .)
NB 5: 81 + 27 = 243 x (. . . .)

Each row of numbers forms a series. Solve for the next/missing number that logically follows:
NB 8: 2, 6, 18, 54, (. . . .)
NB 9: 256, 192, 144, 108, (. . . .)
NB 10: 1, 3, 7, 15, (. . . .)
NB 11: 6, 13, 20, 27, (. . . .)
NB 12: 49, 64, 81, 100, (. . . .)

In each line below the three numbers on the left are related in the same way as the three numbers should be on the right. Solve for the missing middle number on the right.
NB 15: 4 (11) 7 : : 8 (. . . .) 5
NB 16: 3 (12) 4 : : 2 (. . . .) 5
NB 17: 661 (122) 295 : : 514 (. . . .) 121
NB 18: 205 (111) 239 : : 176 (. . . .) 124
NB 19: 784 (112) 336 : : 968 (. . . .) 363

Solve for the missing number that belongs at that step in the series.
NB 22: 52, 45, (. . . .), 31
NB 23: 43, 35, (. . . .), 19
NB 24: 416, 390, (. . . .), 338
NB 25: 92, 79, (. . . .), 53
NB 26: 1, 5, 13, (. . . .), 61

References:

1 Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
2 Serebriakoff, V. *Self-Scoring IQ Tests*. (Sterling Publishing Company, Incorporated, 1996).
3 Zarifhonarvar, A. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925* (2023).
4 Greshake, K. *et al.* More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *arXiv preprint arXiv:2302.12173* (2023).