How people misinterpret answers from Large Language Models

Yuzhi Pan and Chris Baber

University of Birmingham, UK

SUMMARY

We presented probability problems to two Large Language Models (LLMs) and asked human judges to evaluate the correctness of the outputs. Neither LLM achieved 100% on the questions but participants did not always spot the errors these made. Two types of human error were identified: i. the LLM answer is correct, but the participant thought it was wrong (especially with the smaller LLM); ii. the LLM answer was wrong, but participants thought it was correct (especially with the larger LLM). Participants tended to trust the LLM when they were unsure how to answer a question and the LLM provided an answer that seemed reasonable and coherent (even if it is actually wrong)

KEYWORDS

Large Language Models; Prompt Engineering; Human Error; Probability.

Introduction

Large Language Models (LLM) have become commonplace. These use Generative Artificial Intelligence, based on 'transformers', to produce answers to questions on the basis of scouring huge repositories of documents. However, LLMs might produce answers which are incorrect. To cope with problems that LLMs have with reasoning tasks (such as mathematics or programming), there has been a growth of interest in 'chain-of-thought engineering' or 'prompt engineering' (Nye et al., 2021; Wei et al., 2022; Chen et al., 2023) in which a user will provide additional the questions, or prompts, for the LLM to produce a more acceptable answer. For example, a user could ask the LLM to produce a solution to a problem in the form of a sequence of steps need to solve that problem. However, there is evidence that 'prompt engineering' approaches are reliant on the size of the LLM, in terms of the number of parameters used, and billion parameter models tend to respond best to these approaches (Wei et al., 2022). As implied above, the user needs to recognise when the answer is 'acceptable' which might require either knowledge or effort on the part of the user. In a study evaluating the use of LLM for fact-checking, Si et al. (2023) note that "Users reading LLM explanations are significantly more efficient than using search engines with similar accuracy. However, they tend to over-rely on the LLMs when the explanation is wrong."

In this paper, we are interested in using LLMs to respond to problems that are not normally part of their standard remit, i.e., mathematical problems relating to probability. Even if they are capable of producing answers which are mathematically correct, LLMs might not be performing mathematical calculation so much as finding instances of similar results in their search of massive datasets. Unless explicitly instructed (either by the human or from the rules it is applying) the LLM is seeking to produce the next most likely word in a sequence, rather than attempting any form of calculation. A recent development is the specialized LLM that can handle mathematical problems, LLEEMMA, which was trained on a corpus of materials containing to mathematical content (Azerbayev et al., 2023). This suggests that there is every reason to expect the mathematical capability of LLMs to improve massively in the coming years (much as the ability of LLMs to generate computer code has improved in the past few years). But it remains a moot point as to whether these will be perfect in all circumstances (Bender et al., 2021). In this study, we wanted to see whether (even for simple probability problems) people could use Prompt Engineering to produce an answer that they believed was acceptable and whether they could when LLMs made errors.

Even at its most basic, probabilistic reasoning can lead to counterintuitive conclusions (Batanero et al., 2005; Fischbein Schnarch, 1997). A common failing is that people can be insensitive to base-rates of phenomena. As an example of this, Kahneman and Tversky (1973) presented participants with 100 personality descriptions of people and asked them to decide if the description was of an engineer or judge. Participants were told that there were 70 of one profession and 30 of the other in each set of descriptions. Knowing this base-rate, there should have been differences in classification (with more 'engineer' being classified from the set with 70 engineers, and vice versa). However, the distributions of classification were the same across both sets which suggests that participants ignored this base-rate data. Given that people might struggle with calculations involving probability, there might be a temptation to turn to automated support, such as a Large Language Model, to provide assistance in solving probability problems.

Method

We took 10 probability problems, suitable for High School pupils, from mathematics revision guides and presented these to two LLMs (the small LLM - Vicuna-33B - has 33 billion parameters, and the large LLM - ChatGPT3.5 - has 175 billion parameters). We accessed both LLMs through https://chat.lmsys.org. The full set of questions is given in Appendix 1 (the question number is used to label the graphs in the results section).

Participants

16 participants (with different levels of knowledge and confidence in working with probabilities) were randomly allocated to two groups. Each group interacted with one of the LLMs.

Procedure

Participants were asked to present each question to the LLM, evaluate its answer and, if necessary, ask additional questions as a means of 'prompt engineering' to encourage the LLM to produce a correct answer. The answers from each LLM were presented to participants who were asked the following questions:

- Do you think the answer is correct or incorrect?
- Why do you think the answer is correct or incorrect?
- How did the chatbot produce this answer?
- What can you do to get the chatbot to change its answer?

In addition to answering these questions, we asked participants to complete a NASA-TLX to explore whether there were differences in perceived workload between the two LLMs.

Results

How well did the LLMs solve the problems?

Using a two-way ANOVA (LLM x question) there was a significant main effect of LLM [F, 187) = 83.9, $p<0.001$], with ChatGPT3.5 having superior performance to Vicuna-33b, and question [F(9,187) = 15.9, $p<0.000$] with some questions being less likely to be answered correctly (figure 1). Even with its superior performance, ChatGPT3.5 only answered 5 questions correctly on 100% of occasion (and Vicuna-33b had no question that it answered correctly on all occasions). This meant a significant interaction of LLM and question [F9,187) = 3.8, $p<0.05$].

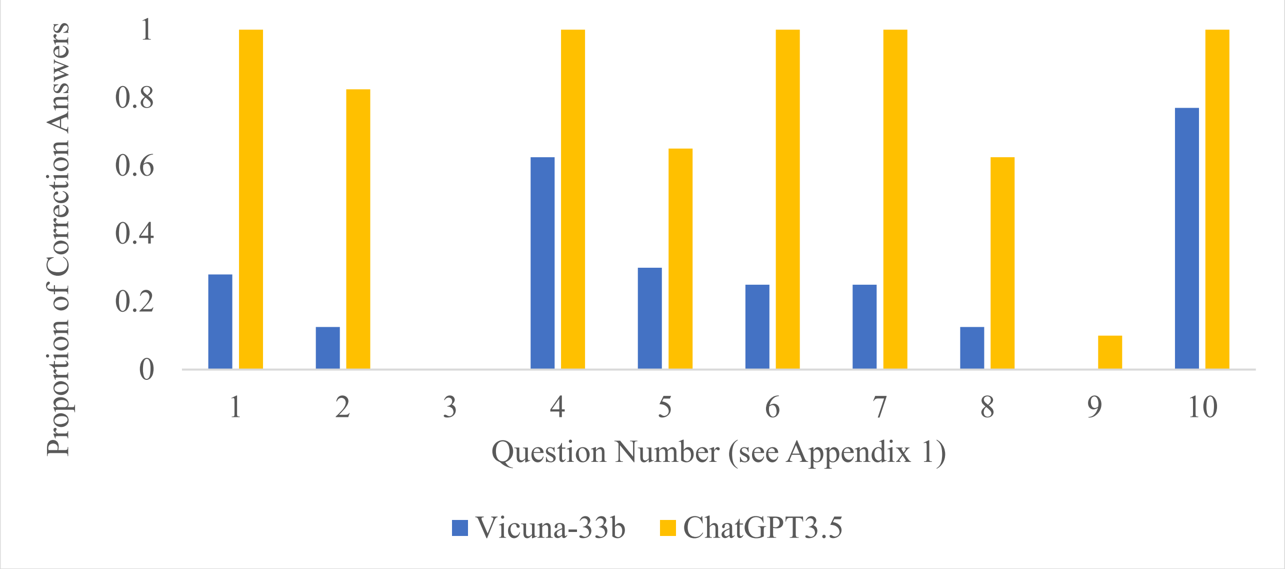

Figure 1: Performance of the LLMs on answering probability questions

Table 1 gives an example of the answers provided by Vicuna-33b and ChatGPT for question 3. One way to solve question 3 is as follows: A customer will report an allergy on 0.08 occasions, so the probability of *not* reporting an allergy is 1 – 0.08 = 0.92. Over the 12 customers in our sample, the probability that none of them will report an allergy is 0.92^{12} = 0.3677 (probabilities are multiplied as move horizontally along the probability tree). So, the probability that at a customer (from this 12) will report an allergy is 1 – 0.3677 = 0.6232, or 63.23%.

Table 1 shows that both LLMs report a method that appears acceptable, but the actual answer was incorrect. If we prompt the LLM with 'the answer is not correct, try again' we get still a range of answers. This illustrates the problems that LLMs have with basic mathematics. As an aside, I asked the same problem of ChatGPT4 and it calculated 0.92^12 as 0.4427, 0.4208, and 0.3770.

Table 1: Examples of Solutions presented by LLMs

Vicuna-33b	ChatGPT3.5
To find the probability that at least one customer will report a food allergy, we will use complementary probability. The complementary probability is the probability that no customers will report a food allergy, which is the opposite of the event we are interested in. The probability that no customers will report a food allergy is (1 - 0.08)^12. Now, we will use the complementary probability rule, which states that the probability of an event occurring is equal to 1 minus the probability of the complementary event occurring. P(at least one customer will report a food allergy) = 1 - P(no customers will report a food allergy) P(at least one customer will report a food allergy) = 1 - (1 - 0.08)^12 P(at least one customer will report a food allergy) ≈ 1 - (0.92)^12 P(at least one customer will report a food allergy) ≈ 1 - 0.394	To find the probability that at least one customer will report a food allergy, we can find the probability that none of the customers report a food allergy and subtract it from 1. The probability that a customer does not report a food allergy is 1 - 0.08 = 0.92. Since the customers are independent and the probability of each customer not reporting a food allergy is the same, we can calculate the probability that none of the 12 customers report a food allergy as (0.92)^12 ≈ 0.428. Therefore, the probability that at least one customer will report a food allergy is 1 - 0.428 = 0.572 or 57.2%.

P(at least one customer will report a food allergy) ≈ 0.606 So, the probability that at least one customer will report a food allergy is approximately 60.6%, rounded to the nearest hundredth.	

Thus, while the LLM produces an incorrect numerical solution, it often presents a plausible process by which it reaches that solution. Therefore, we might expect participants to agree with an incorrect answer because the process looks correct.

How often did participants Misunderstand the LLM output?

We define misunderstanding LLM output as participants accepting an answer when it is incorrect or rejecting an answer when it is correct. Figure 2 shows the proportion of participants who made an error in accepting (or rejecting) the LLM solutions.

As figure 2 illustrates, there was no main effect of LLM on this measure but there was a significant main effect of question [F(9,211) = 4.2, p<0.0001] and a significant interaction [F(1,211) = 2.1, p<0.05].

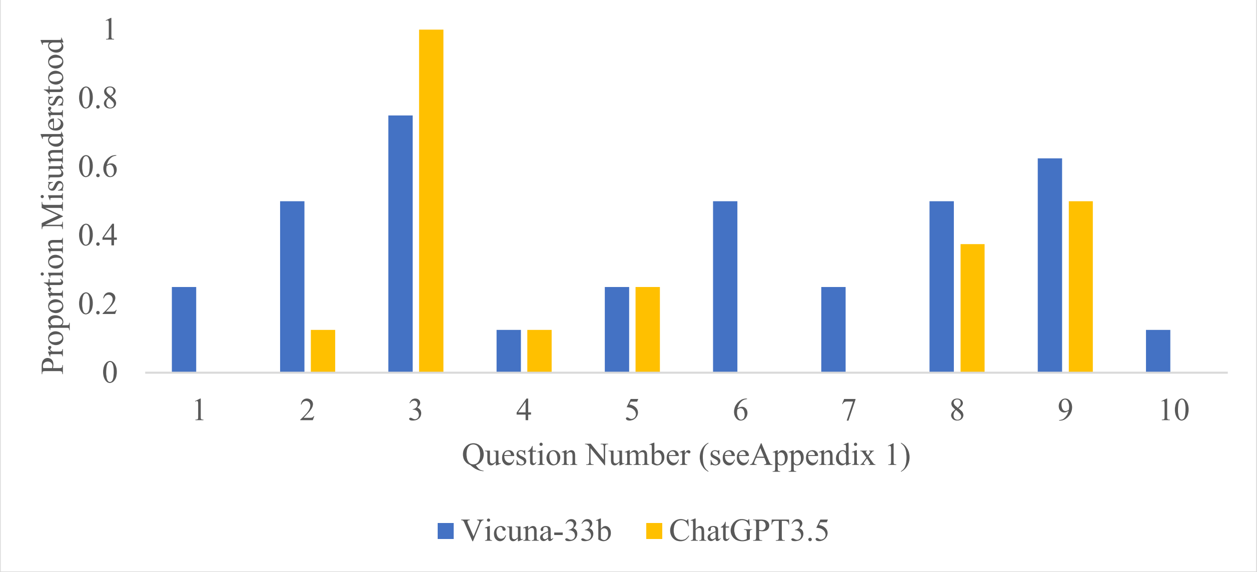

Figure 2: Proportion of Participants who Misunderstood the LLM answer

For question 3 which, we noted above, involves straightforward problem-solving steps but complex calculations, participants tend to trust the LLM answers. This is because the LLM explains a process which appears clear and logical, even though its calculations might be challenging for participants to verify. Question 9 was challenging for several participants and, also in this case, the clear and logical approach of each LLM, led participants trust the responses. Table 2 gives some examples of comments that participants made when reviewing the output from each LLM.

Table2: Examples of comments

	Question 3	Question 9
Vicuna-33b	The approach is the same as mine, but there seems to be a calculation error. (P1) The procedures are logical, but the answer is approximately 1, which seems a bit odd. (P2)	Misunderstanding the question led to using incorrect data for calculations. (P4) The chatbot's problem-solving approach is the same as mine. It extracts numbers and calculates

		the answer based on the algorithm. (P8)
ChatGPT3.5	The response contains a detailed textual explanation and doesn't include a formula; instead, it directly uses numerical values for calculation. The calculation process is accurate. (P10) The calculation steps are the same as my approach (P16)	The responses of the chatbot exhibit logical issues, and there are errors in incorporating the data, resulting in the generation of clearly incorrect answers "0”. (P9) Number of people who like pepperoni pizza (B) i don't think it is 50 (P14)

Is there a relationship between Knowledge of Probability and Misunderstanding?
The participants provided information on the level to which they studied probability in mathematics (or related) classes. 1 participant only studied this to junior school, 3 studied it to high school, 5 to Undergraduate degree, and 7 to Postgraduate degree level. The highest proportion of misunderstand comes from participants who had studied probability to Undergraduate level (5 / 50 (i.e., 5 people x 10 questions each), or 10%. We also asked participants how often they used concepts from probability: Never = 5; at least one degree module = 6; part of my research or project work = 5. Again, participants who were studying probability for at least one degree module showed the highest frequency of misunderstanding (4 / 60 = 6.7%). We asked participants to rate their confidence in using probability: Not at all confident = 1; Slightly confident = 5; Moderately confident = 8; Very confident = 2. In this case, misunderstanding appears to relate to slightly or very confident (as shown in figure 3).

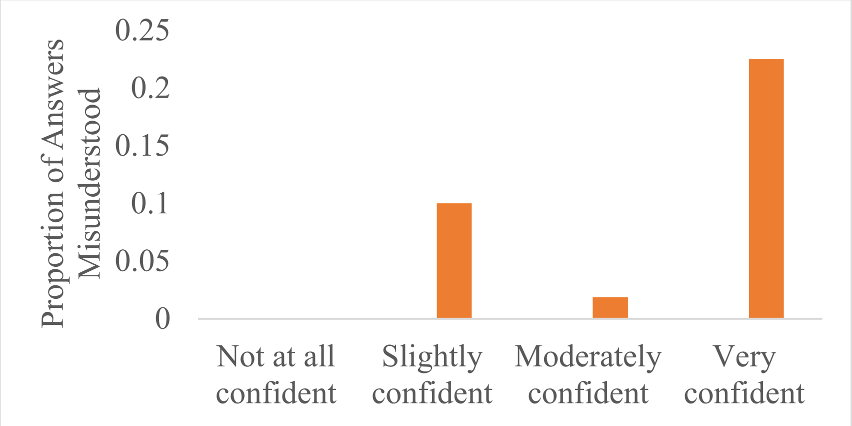

Figure 3: Proportion of Answer Misunderstand for self-rated Confidence in using Probability

Does subjective workload vary with LLM complexity?
Figure 4 shows that workload tends to be higher for the smaller LLM. This might be because the smaller LLM made more errors and participants expected to have to work harder to check its answers. Or it might be that the put less effort into working the larger LLM because they trusted its answers.

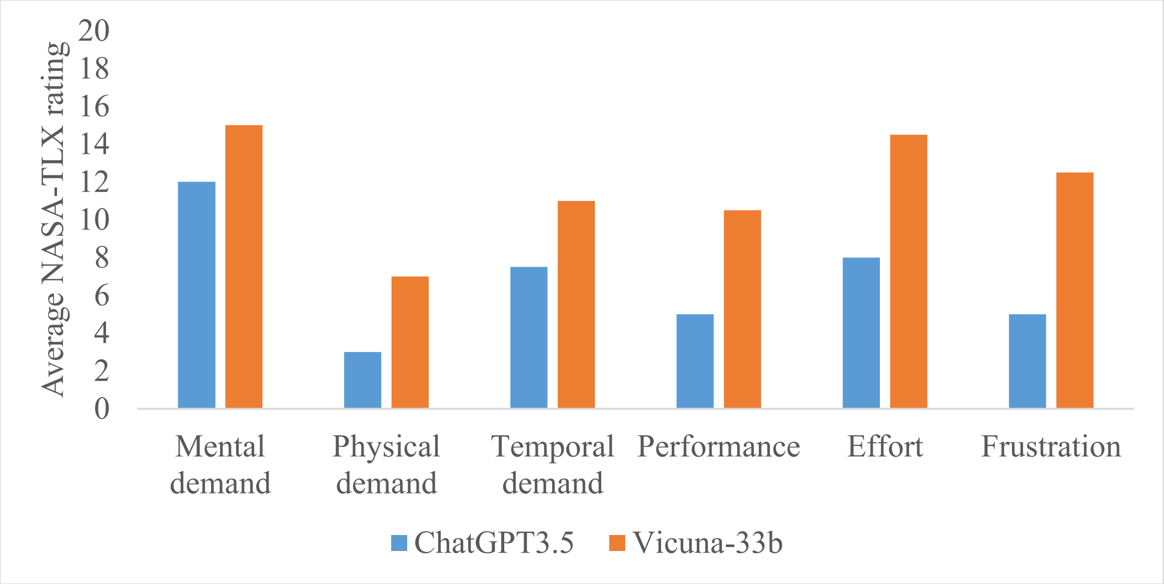

Figure 4: Workload Ratings from NASA-TLX

Conclusions

Both LLMs chatbot does perform poorly on math problems. Possible reasons include:

- Model training data issues: If an LLM rarely encounters relevant questions in its training data, or if those questions are expressed in a variety of ways, it may have difficulty parsing such questions.
- Limitations of information extraction algorithms: Overly simple keyword extraction or sentence parsing means may be difficult to capture all the details of the problem.
- Difficulty in context analysis: LLMs may have errors in interpreting the context and context of a sentence.
- Computing power limitations: In order to achieve rapid response, some real-time systems may sacrifice in-depth analysis steps, thereby missing critical details.

ChatGPT3.5 was more accurate than Vicuna-33b in answering probability questions. But that does not mean ChatGPT3.5 was flawless. In fact, for some questions, its answering level is not much different from Vicuna-33b's, and for some questions, it does not give the correct answer. It is worth noting that the questions with high error rates often involve complex calculations and require multi-step solutions.

Compared with ChatGPT3.5, Vicuna-33b's ability to answer probability questions is obviously lacking. Vicuna-33b was not able to interpret some the questions, such as those involving summation. For example, it did not identify the values to be added, or it made the wrong reduction after calculating the correct fraction. Vicuna-33b also failed to provide correct answers even when users clearly pointed out the error and gave corrections, such as "4/25 is not equal to 1/6" or "4/25=0.16" reply.

When the LLMs made mistakes, these were not always spotted by participants. Misunderstandings can be divided into two main categories:

- The Chatbot's answer is correct, but the participant thinks it is wrong: This situation is more common in Vicuna, especially when participants are unsure about or do not understand the question. Some misunderstandings stem from participants' own biased understanding of the topic, while other misunderstandings stem from their distrust of Vicuna. For example, some participants expressed distrust in Vicuna because of simple calculation errors in the first few questions.
- The chatbot's answer is wrong, but the participant thinks it is correct: While this situation exists in both chatbots, it is more common in ChatGPT. Part of the reason is that participants have excessive trust in ChatGPT, which may stem from their previous understanding of ChatGPT or their lack of confidence in solving probability problems. The answer to question three is in decimal form and is very close to the correct answer, making it difficult for participants to distinguish. More commonly, participants tend to trust the chatbot when they are unsure of how to answer a question and the chatbot provides an answer that seems reasonable and coherent (even if it is actually wrong).

The task load index, it was found that Vicuna's interactions significantly increased the burden on participants which may be related to the lower accuracy rate. Vicuna often takes a more sophisticated approach when answering questions. For example, even though a simple formula could have been used, it chose the enumeration method, which not only made the answer lengthy but also more likely to cause an error warning on the page. In addition, the page response time of VICUNA is significantly longer than that of ChatGPT, which further increases the pressure on participants. These factors may also have affected participants' understanding and misunderstanding of robots to a certain extent.

Discussion

Exploring misunderstandings of the LLM outputs in more detail, we found that knowledgeable users are over-confident in their abilities and did not check the output from the LLMs, i.e., their errors might be due to a lapse of attention in checking the output. There are many factors that contribute to human misunderstanding of LLMs. According to this study's results, combined with qualitative interview data, if the question is challenging for the participant or involves complex calculations, and neither the LLM nor humans can give accurate answers, the detailed steps provided by the LLM may lead to misunderstandings by participants. In addition, some external factors, such as previous experience with LLMs and participants' confidence in understanding probability, may also affect their trust and perception of both LLMs.

References

Azerbayev, Z., Schoelkopf, H., Paster, K., Dos Santos, M., McAleer, S., Jiang, A.Q., Deng, J., Biderman, S. and Welleck, S., 2023, Llemma: an open language model for mathematics, *arXiv: 2310.10631.*

Batanero, C., Henry, M., and Parzysz, B., 2005, The nature of chance and probability, In *Exploring probability in school: Challenges for teaching and learning*, Springer, 15-37.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Chen, B., Zhang, Z., Langrené, N. and Zhu, S., 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review, *arXiv:2310.14735.*

Fischbein, E. and Schnarch, D., 1997, Brief report: The evolution with age of probabilistic, intuitively based misconceptions, *Journal for research in mathematics education, 28*, 96-105.

Nye, M., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D. and Sutton, C., 2021, Show your work: Scratchpads for intermediate computation with language models, *arXiv:2112.00114.*

Si, C., Goyal, N., Wu, S.T., Zhao, C., Feng, S., Daumé III, H. and Boyd-Graber, J., 2023. Large Language Models Help Humans Verify Truthfulness--Except When They Are Convincingly Wrong. arXiv preprint arXiv:2310.12558.

Tversky, A., and Kahneman, D., 1973, Availability: A heuristic for judging frequency and probability, *Cognitive Psychology*, *5*, 207-232.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, pp.24824-24837.

Appendix 1: Set of Probability Problems used in the Study

1. David is at a car dealership. He is going to randomly select a vehicle to test drive. There are 13 trucks, 8 vans, and 4 compact cars. What is P(compact car)?
2. A jar contains 4 red marbles, 4 green marbles, and 5 blue marbles. If we choose a marble, then another marble without putting the first one back in the jar, what is the probability that the first marble will be blue and the second will be green?
3. When a customer places an order at Ying Ying's bakery, there is an 8 % probability that the customer will report a food allergy. One day, 12 customers placed orders at Ying Ying's bakery. Assuming that each of the 12 customers is equally likely to report a food allergy, what is the probability that at least one customer will report a food allergy? Round your answer to the nearest hundredth.
4. Mary surveyed a random sample of students in her school district, and she found these statistics: P(rides bike to school)=0.14; P(has crossing guard)=0.48; P(rides bike and crossing guard)=0.12. Find the probability that a student rides a bike to school, given that the student's school has a crossing guard.
5. You're playing a game where you defend your village from an orc invasion. There are 3 characters (elf, hobbit, or human) and 5 defense tools (magic, sword, shield, slingshot, or umbrella) to pick from. If you randomly choose your character and tool, what is the probability that you won't be a hobbit or use an umbrella?
6. Stephen read 12 books, 20 magazines, and 17 newspaper articles last year. Based on this data, what is a reasonable estimate of the probability that Stephen's next reading material is a magazine? Choose the best answer. a. 1、0.20;2、0.24;3、0.35;4、0.41
7. Captain Jessica has a ship, the H.M.S. Khan. The ship is two furlongs from the dread pirate Luis and his merciless band of thieves. The Captain has probability 4/9 of hitting the pirate ship. The pirate only has one good eye, so he hits the Captain's ship with probability 1/ 4. If both fire their cannons at the same time, what is the probability that both the pirate and the Captain hit each other's ships?
8. If you flip three fair coins, what is the probability that you'll get at least two heads?
9. Sam asked 50 people whether they like vegetable pizza or pepperoni pizza.
   37 people like vegetable pizza.25 people like both 3 people like neither.
   Sam picked one of the 50 people at random. Given that the person he chose likes pepperoni pizza, find the probability that they don’t like vegetable pizza.
10. The probability that a biased dice will land on a five is 0. 4. Lewis is going to roll the dice 400 times. Work out an estimate for the number of times the dice will land on a five.