Additional File 2:

Categorisation of answer

Examples of categorization of correct answers

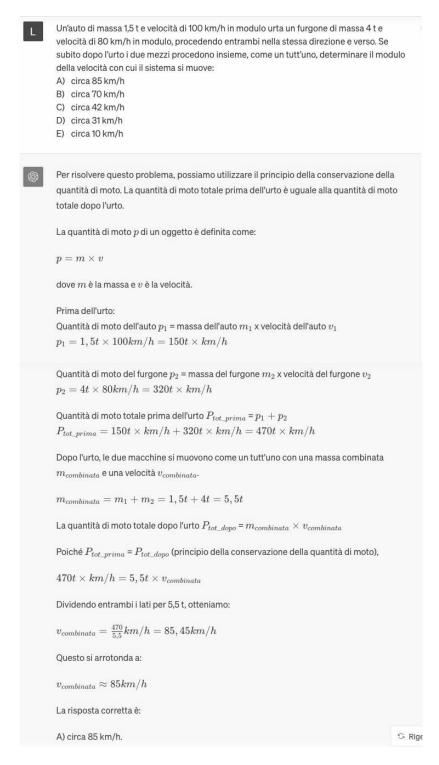


Figure 1a: Example of a correct answer labelled as "Logical reasoning", as the chatbot (ChatGPT-4) clearly demonstrate the logic presented in the response (question n. 51, CINECA test - 2023).

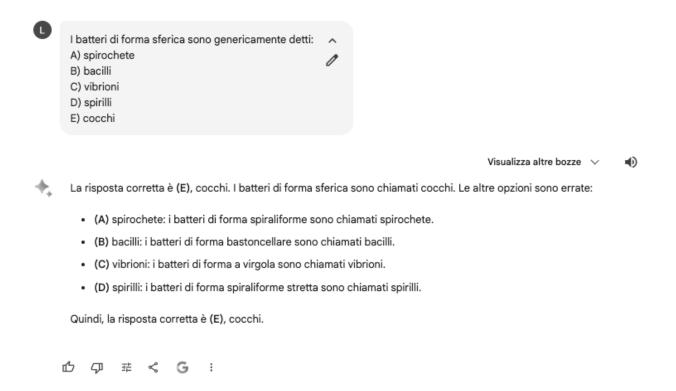


Figure 1b: Example of a correct answer labelled as "Internal information", as the chatbot (Bard) includes information internal to the question (question n. 29, CINECA test - 2019).

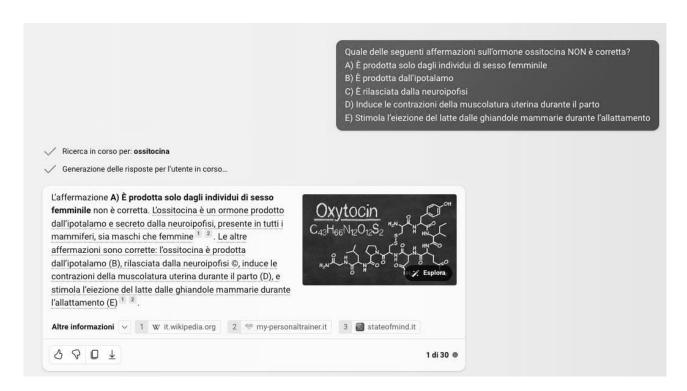


Figure 1c: Example of a correct answer labelled as "External information", as the chatbot (Microsoft Bing) use information external to the question to provide the answer, performing a web search and indicating the origin of the source from which the information is taken (question n. 30, CINECA test - 2023).

Examples of categorization of incorrect answers



Figure 2a: Example of an incorrect answer labelled as "Logical error", as the chatbot (Microsoft Bing) correctly identifies the relevant information but fail to convert it into an appropriate answer (question n. 10, CINECA test - 2016; correct answer: E).

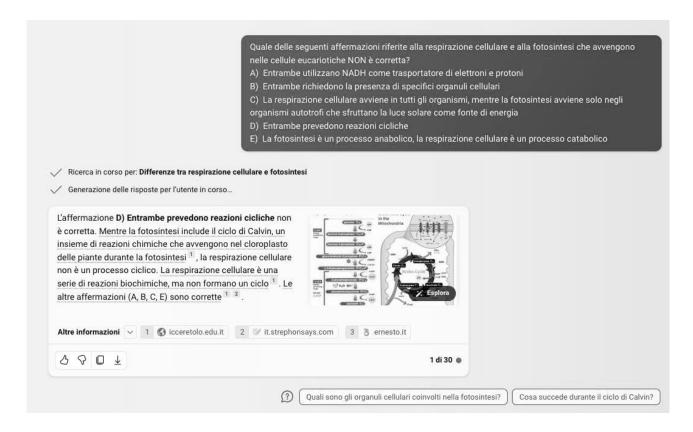


Figure 2b: Example of an incorrect answer labelled as "Information error", as the chatbot (Microsoft Bing) fails to recognize a key piece of information, whether present in the question stem or through external information (question n. 20, CINECA test - 2023; correct answer: A).

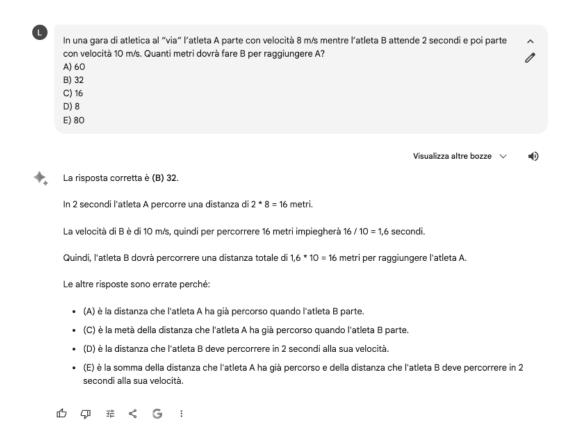


Figure 2c: Example of an incorrect answer labelled as "Statistical error", as the chatbot (Bard) commits an error that involves an arithmetic mistake (question n. 76, CINECA test - 2011; correct answer: E).