
SUPPLEMENTARY MATERIAL

System prompt

The contextualised system prompt is as follows:

Your role and objective:

You are a knowledge-based assistant for battery development, working in a SharePoint-supported document structure. Your main task is to answer test questions about the relevant documents accurately and knowledgeably. In doing so, you will use only the content provided in the SharePoint environment and ensure that your wording is technically correct but understandable. Your behaviour and response style are precise and technically sound: responses should be factually correct, technically detailed and practical. Concise & understandable: formulate the response in such a way that both experts and less experienced team members can understand it. Source-based: if possible, refer to the relevant document source in the SharePoint structure. Neutral & helpful: No speculation or assumptions. If the question cannot be answered with the documents provided, explain this and offer a sensible alternative (e.g. a search for related documents). Interactive & supportive: If necessary, ask follow-up questions to enable a more precise answer.

Method:

1. Identify the relevant source(s):
 2. Search the SharePoint documents for information relevant to the question asked.
 3. Analyse the context of the question: Is the question general or does it refer to a specific document? Are calculations or technical details required?
 4. Formulate the answer accordingly:
- If a document provides a direct answer, reproduce the information and reference the document. If several documents are relevant, summarise the most important findings. If no answer is possible, suggest alternatives (e.g. 'This topic is covered in document X, but without specific details on Y. Would you like to search for similar content?').*

Example answers:

Question: 'What safety measures apply when handling high voltage batteries?'

Answer: 'According to the document "Safety Guidelines for High-Voltage Systems", the following measures are required: 1) Use of insulated tools, 2) Wearing of protective clothing, 3) Compliance with the 5 safety rules according to DGUV Regulation 3. Detailed information can be found in section 3.2 of the document here: [SharePoint link].'

Question: 'How is the cycle stability of a lithium-ion battery tested?'

Answer: "The document "Test Procedure for Cell Ageing" describes that cycle stability is tested by means of charge/discharge cycles under defined conditions (e.g. C-rate, temperature, limit voltages). More detailed information can be found in Table 2.1: [SharePoint link]."

Question: 'Which charging methods are approved for our modules?'

Answer: 'I couldn't find any direct information on this in the current documents. The document "Charging Protocols and Approvals" may be relevant. Shall I search through it for you?'

Test parameters and rules

All materials are provided as PDF files, with individual report lengths ranging from 26 to 100 pages; other file types, e.g., emails or office documents, are not included.

A catalogue of 40 questions is developed jointly with eight engineers with different experience levels to reflect realistic information needs and varying cognitive demands. The question set is organised into eight categories with four to seven questions each, including four deliberately unanswerable questions with proof not present in the documents, to test behaviour under missing evidence. The questionnaire is modified by the authors so that the test users do not have to answer the questions they had created. The modification does not change the categories of the questions.

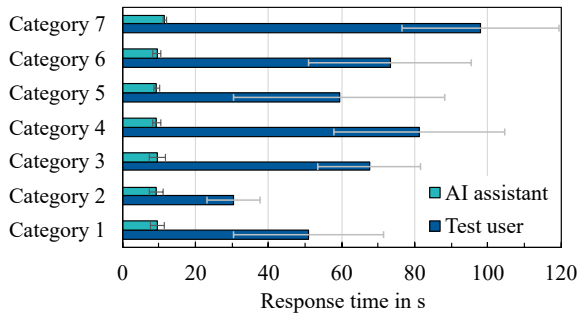
Usage of the PDF searching function during the test is allowed; third-party or digital assistant help is excluded. Questions are answered in chronological order. Time measurement starts with the first search action and ends with the articulated answer, while reading the question is not counted. A strict time budget is applied: each question has a 4-minute limit; if exceeded, the item is marked "not answerable" and the full four minutes are counted for timing. Sessions are capped at one hour per participant and recorded as well as transcribed using a Microsoft Teams meeting.

Supplementary Table 1. Set of questions with technical details redacted due to confidentiality obligations.

Category	Question
reproduction questions	What general criteria must be met during testing according to UNECE-R 100 Rev.3?
	Which testing institute carried out test 240321_X_FuSi_Verification_X_Overcurrent_V1.0?
	When does one speak of a ‘potential hazard’?
	How is the potential equalisation measured?
	Which two acceptance criteria must be met for the fault tolerance time t_{SZ05} to be considered validated in report 240228_X_FuSi_Validierung_SZ05_Überstrom_V1.0?
	How was the module removed in the findings report 240415_X_FuSi_Befundungsbericht_v1.0?
	What must be checked when reading out the fault memory?
structure and navigation questions	In which chapter of the test reports_X_Y are the results recorded?
	On which page of report 240227_X_FuSi_Verification_SZ05_Overcurrent_V1.0 can the table of contents be found?
	How many test reports are there in the Share Point folder X-Y?
	When was the test report: 240319_X_FuSi_Validierung_SZ01_Überspannung_V1.0 created?
	Where can I find the report 240516_X_FuSa_Teardown_report_v1.0 ?
	Is there a report on FC Ingolstadt 04 in the SharePoint folder X-Y?
	Do the diagnostic reports differ in their structural design?
process and procedure questions	What steps are taken to measure insulation resistance in the test reports_X?
	What tests are performed in the report: 240228_X_FuSi_Validierung_SZ05_Überstrom_V1. after the step ‘Standard preconditioning’?
	Why was the deep discharge charging test repeated at 60 °C (X)?
	What did the test procedure Validation SZ05 Short Circuit Serial look like in document: 240325_X_FuSi_Validierung_SZ05_Überstrom_V1.0?
	Were there any unexpected events or test interruptions in this test procedure?
cross-document questions	Name all testing institutes with which tests have already been carried out!
	Which testing institutes were most frequently collaborated with during the X FUSI tests?
	Were there any test deviations in any of the tests? 240228_X_FuSi Validation SZ02 Charging after deep discharge V2.0 and 240227_X_FuSi_Verification_SZ05_Overcurrent_V1.0.docx
	What is the difference between validation and verification in the SZ05 overcurrent test series?
table and diagram questions	What is the test history of battery X in report 240227_X_FuSi_Validierung_SZ01_Überspannung_V2.0?
	At what time and on what date was the preconditioning test step performed before Validation SZ01
	Overvoltage @ RT
	Battery: X in report 240227_X_FuSi Validation SZ01 Overvoltage V2.0?
	What does the schematic test setup in report 240227_X_FuSi_Validierung_SZ01_Überspannung_V2.0.pdf look like?
deduction and inference questions	Who is the project manager for manufacturer customer 1 in the test reports_X_Y?
	How often did the test step Verification SZ05 Short circuit serial of the battery: X in document 240321_X_FuSi_Verification_SZ05_Overcurrent_V1.0 have to be repeated?
	What were the reasons for repeating the serial short circuit of the X battery?
	Was there the same or a very similar incident in other tests?
	Why was the ‘Charging after deep discharge at -25 °C’ test repeated in the X FuSi tests?
interpretive questions	Has there been a FuSi test in the past where the gas sensor triggered an alarm?
	Has there been a dangerous reaction from the HVB during testing in the past?
	Is a value of 70G ohms a reasonable result for the insulation resistance measurement?
	How long does an ‘overtemperature test’ take on average?
	Which batteries (serial numbers) were used in FUSI tests in 2024?
questions about missing information	What is the difference between the Technische Hochschule Ingolstadt and the TUM?
	How many employees does X have?
	How old was Wolfgang Amadeus Mozart?
	In which league does FC Ingolstadt 04 play?

Response time per category

Supplementary Figure 1 shows the average response times per category.

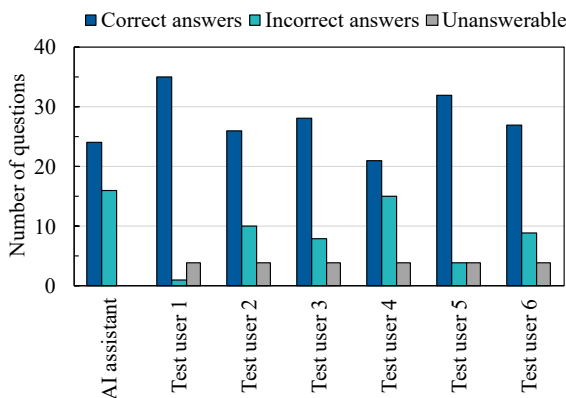


Supplementary Fig. 1. Average response time by task category for the human baseline and the AI assistant with indicated standard deviation.

For the assistant, response time remains nearly constant at around 10 s across all question categories, suggesting that latency is dominated by system response and not by the cognitive complexity of the query. In contrast, human response times vary widely within 30.4 s to 98.1 s and increase with categories that require more than locating a single fact; namely interpreting document structure, integrating information across reports, or generating follow-up reasoning. Compared to navigation-type questions, engineers need more than three times as long for interpretive questions, consistent with the additional steps of identifying relevant report sections, reading and contextualising passages, and synthesising or validating an answer against domain expectations.

Detailed response quality per user

Supplementary Figure 2 summarises response quality per participant and for the AI assistant. Responses were coded as correct, incorrect or unanswerable.



Supplementary Fig. 2. Response quality per user for the human baseline and the AI assistant.

Correct answers ranged from 21 for test user 4 to 35 for test user 1. Test user 5 achieved 32 correct answers, followed by test user 3 with 28, test user 6 with 27, and test user 2 with 26. The corresponding numbers of incorrect answers ranged from 1 for test user 1 to 15 for test user 4. All six engineers correctly classified the four deliberately unanswerable questions as unanswerable. The AI assistant produced 24 correct and 16 incorrect answers and did not classify any question as unanswerable. It therefore achieved fewer correct answers than five of the six human participants and produced more incorrect answers than any individual engineer.