

PowerGLUE: A comprehensive benchmark for evaluating large language models in the electric power domain

Cuiqi Zheng

CSG Digital Grid Group Co., Ltd

Chengwei Peng

CSG Digital Grid Group Co., Ltd

Shouyu Liang

lsynd@126.com

CSG Digital Grid Group Co., Ltd

Zhuojun Cai

CSG Digital Grid Group Co., Ltd

Zhichen Peng

CSG Digital Grid Group Co., Ltd

Zehong Chen

CSG Digital Grid Group Co., Ltd

Article

Keywords:

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11029574/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Existing evaluation benchmarks for large language models (LLMs) do not adequately capture the professional knowledge, normative constraints, and engineering calculations that characterize the electric power sector. Here we introduce PowerGLUE, a domain-specific benchmark for evaluating LLMs in this sector. PowerGLUE covers the six core business sectors of power systems, namely generation, transmission, transformation, distribution, dispatching, and electricity consumption, and consists of two complementary task categories, knowledge question answering and reasoning calculation, organized into seven subtasks. It is constructed from authoritative sources, including technical standards, registered electrical engineer examination questions, textbooks, and engineering cases, followed by a multi-stage quality-control process that combines automated verification and expert review; the current version contains 3,603 items. We evaluate seven representative LLMs and find substantial performance differences across models, task types, and business sectors. Under the current item composition and scoring protocol, exact-match accuracy is generally higher on the reasoning-calculation track than on the knowledge-question-answering track, a pattern that suggests measurable differences between the two tracks, although differences in item difficulty and composition preclude direct causal attribution to knowledge or reasoning ability. PowerGLUE provides a systematic framework for examining the capabilities and boundaries of LLMs in safety-critical power applications.

Introduction

Electric power systems are characterized by strong professional specialization, numerous operating constraints, and stringent safety requirements. As the construction of new power systems dominated by renewable energy advances, grid operating modes become increasingly complex, and the coordination of source, grid, load, and storage faces unprecedented challenges [1]. The industry's demand for capabilities such as intelligent question answering, regulation interpretation, fault analysis, and assisted calculation keeps growing. In recent years, large language models (LLMs) have shown strong capabilities in natural language understanding and complex reasoning, opening a new technical route to intelligent applications in power scenarios [2, 3].

The power domain, however, is not a simple extension of general-purpose language tasks: it involves a large body of specialized terminology, standard clauses, equipment principles, and engineering calculations, and places high demands on the accuracy, consistency, and traceability of results, where factual errors, misquoted regulations, or computational deviations may undermine the reliability of engineering decisions [4]. Before LLMs are introduced into power scenarios, it is therefore necessary to establish a dedicated evaluation benchmark covering the depth of industry knowledge, normative constraints, and engineering reasoning [5].

Existing evaluation systems fall short of these needs. General-purpose benchmarks such as GLUE [6], SuperGLUE [7], MMLU [8], and C-Eval [9] mainly target general language understanding or interdisciplinary knowledge, and can hardly reflect the power domain's requirements for understanding standards and regulations, grasping equipment mechanisms, and solving engineering problems (Table 1). In high-threshold vertical domains such as medicine (e.g., MedQA [10]), finance (e.g., FinEval [11]), and architecture (e.g., Arch-Eval [12]), building dedicated benchmarks has become a standard paradigm for advancing deployment [13], and preliminary efforts have emerged in the power domain itself [14, 15].

Benchmarks for LLMs have evolved from general language understanding toward multi-disciplinary comprehensive assessment. As shown in Table 1, early works such as GLUE and SuperGLUE established the evaluation paradigm for natural language processing around inference, matching, and reading tasks, and MMLU, MMLU-Pro [16], and C-Eval subsequently extended evaluation to multi-disciplinary knowledge question answering, shifting the focus from single language abilities toward broader knowledge assessment.

Table 1
Comparison of existing general-purpose LLM evaluation benchmarks.

Benchmark	Power-related knowledge involved	Standards and clauses	Engineering calculation items	No. of items	Item type
MMLU-Pro [16]	Marginal; mainly electricity-related content in university physics	No	No	12,032	Multiple choice
C-Eval [9]	Marginal; includes 381 registered electrical engineer examination questions	No	No	13,948	Multiple choice
GLUE [6]	None; focused on natural language understanding	No	No	≈ 425k (test)	True/false and open questions
SuperGLUE [7]	None; focused on natural language understanding	No	No	≈ 20.3k (test)	True/false and open questions

Although these benchmarks provide an important reference for assessing general capabilities, their task designs and data sources are oriented toward open-domain corpora or broad disciplinary scenarios, and power-related content in them accounts for a low proportion and remains at the level of basic concepts. The limitations are threefold: insufficient depth of power-domain knowledge coverage, which makes it difficult to reflect mastery of the industry knowledge system; scarcity of engineering reasoning and calculation tasks, which makes it difficult to uncover problem-solving ability in real business scenarios; and insufficient attention to industry normativity, business constraints, and result reliability in high-risk professional scenarios. Directly adopting general benchmarks to evaluate the suitability of LLMs in the power domain therefore rarely yields comprehensive and accurate conclusions.

Existing efforts fall roughly into two categories (Table 2). On the vision side, datasets such as InsPLAD [17] and CPLID [18] target defect detection and classification of transmission equipment; they are valuable for equipment-state identification, but their task scope is mostly confined to local inspection scenarios, leaving professional knowledge understanding, fault-mechanism analysis, and response-strategy reasoning largely unexamined. On the text side, two independent lines of work share the name ElecBench; to avoid confusion we refer to them as ElecBench-Dispatch [14] and ElecBench-EAAI [15]: the former constructs a professional evaluation set for dispatch scenarios, probing robustness dimensions such as misinformation and hallucination across general knowledge, monitoring, dispatching, and black-start test sets, while the latter assembles 24 datasets with 34,030 entries covering general power knowledge and four business applications. These works markedly improve industry adaptation, but ElecBench-Dispatch is organized around a single business link, while ElecBench-EAAI, despite its scale, is a collection of heterogeneous datasets without engineering calculation tasks or a unified sector-task-subtask structure.

Table 2
Comparison of existing power-domain LLM evaluation benchmarks.

Benchmark	Power sectors covered	Standards and clauses	Engineering calculation items	No. of items	Item type
PowerGLUE (this work)	All six sectors: generation, transmission, transformation, distribution, consumption, dispatching	Yes	Yes	3,603	Multiple choice
ElecBench-Dispatch [14]	Dispatch-oriented: general, monitoring, dispatch, and blackstart test sets	No	No	288	Open and true/false questions
ElecBench-EAAI [15]	General power knowledge; equipment, dispatching, materials, and marketing	Yes	No	34,030 (24 datasets)	Mixed: QA, generation, classification, extraction
InsPLAD [17]	Transmission	No	No	10,607	Object detection and image classification
CPLID [18]	Transmission	No	No	848	Object detection and image classification

Overall, dedicated power-domain evaluation has progressed from single tasks toward professional-scenario adaptation, but gaps remain in coverage, task-form diversity, and diagnostic granularity: evaluation structured uniformly across the six business sectors is lacking; engineering calculation, fault analysis, and scheme optimization are underrepresented relative to text understanding and visual recognition; and few benchmarks combine sector-, task-, and subtask-level analyses within one framework, which limits their ability to localize model weaknesses. These observations motivate a benchmark oriented to the core business scenarios of the power industry that balances knowledge question answering with reasoning calculation and supports multi-dimensional analysis.

To this end, we propose PowerGLUE, a capability-evaluation benchmark for LLMs in the power industry. The benchmark covers the six core business sectors of generation, transmission, transformation, distribution, dispatching, and consumption, and sets two task categories, knowledge question answering and reasoning calculation, to assess models' professional knowledge understanding, application of standards and regulations, and engineering reasoning and problem solving. The main contributions of this work are threefold. First, we construct an evaluation framework organized around the business chain of power systems, in which each of the six sectors is examined under both task tracks, so that the benchmark reflects the applicability and capability boundaries of models across the industry chain. Second, we design a dual-track task system with seven subtasks that separates knowledge-oriented performance from engineering application, supporting fine-grained diagnosis rather than a single overall score. Third, we build the dataset through a traceable construction pipeline over multi-source professional materials, namely technical standards, registered electrical engineer examination questions, textbooks, and engineering cases, with two-model consistency classification, retrieval-constrained supplementary generation, and expert verification, and we report a comparative evaluation of seven mainstream models. Experimental results show that the capabilities of existing models in professional power scenarios differ markedly across models, sectors, and task types.

Results

The PowerGLUE benchmark

PowerGLUE follows three design principles: core-business coverage, equal emphasis on knowledge and application, and fine-grained diagnosis. (1) Core-business coverage. The power industry has an evident chain-like structure, and performance in one local scenario cannot represent suitability for the industry as a whole; the framework therefore covers generation, transmission, transformation, distribution, dispatching, and consumption, so as to avoid one-sided results caused by an overly narrow task scope. (2) Equal emphasis on knowledge and application. Power-domain tasks include both the understanding of professional concepts, standard clauses, and technical parameters, and application-oriented problems such as fault analysis, numerical calculation, and scheme optimization; the two task categories, knowledge question answering and reasoning calculation, are therefore set in parallel to assess knowledge mastery and engineering application respectively. (3) Fine-grained diagnosis. Beyond overall accuracy, a professional evaluation should capture differences across capability dimensions; as advocated by the HELM holistic evaluation framework, a single overall score easily masks severe defects in specific dimensions [19]. PowerGLUE therefore sets seven subtasks under the two categories, supporting analysis at the levels of terminology understanding, standard interpretation, parameter memorization, equipment principles, fault analysis, engineering calculation, and planning optimization. On this basis, the task system consists of six business sectors, two main task categories, and seven subtasks.

Following the business chain of production, transmission, conversion, distribution, regulation, and consumption of electric power, PowerGLUE divides the evaluation scope into six core sectors: generation (equipment operation, process flows, and operating indicators of electric energy production), transmission (lines, towers, insulation, and corridors), transformation (substation equipment, primary and secondary systems, and operation and maintenance), distribution (network structure, equipment operation, and fault handling), dispatching (grid operation control, dispatching rules, and operating logic), and consumption (customer-side power supply and utilization, energy metering, and related application scenarios). The sectors are delimited by the actual business links of power systems rather than by single devices or single job positions, so that the evaluation content covers fairly typical industry scenarios and provides a unified scenario framework for subsequent task design.

To balance the examination of static knowledge and dynamic problem solving, PowerGLUE adopts a dual-track task system. Knowledge question answering examines memorization, understanding, and discrimination of power-domain professional knowledge, and comprises four subtasks: (1) terminology definition and understanding: the ability to understand power-domain terms, basic concepts, and their scope of application; (2) standard clause interpretation: the ability to understand and discriminate power standards, regulations, and technical clauses; (3) numerical indicators and parameter memorization: the degree of mastery of common technical parameters, operating indicators, and numerical ranges; and (4) power equipment principles: the degree of understanding of the working mechanisms, structural characteristics, and operating principles of typical power equipment. Reasoning calculation examines the ability to analyze, derive, and solve problems in complex business scenarios, and comprises three subtasks: (1) fault diagnosis and analysis: cause analysis and logical judgment on the basis of fault phenomena, operating states, or protection information; (2) power calculation application: solving problems involving formula application, parameter conversion, and engineering calculation; and (3) planning, design, and optimization: comprehensive analysis in problems such as scheme comparison, operating-mode selection, and design optimization. The two tracks complement each other: the former reflects the model's knowledge foundation, and the latter its level of application in engineering scenarios.

Recent theoretical studies of frontier artificial intelligence indicate that autoregressive Transformer architectures have an intrinsic weakness, tending to diverge or stall between steps, when handling multi-step mathematical derivation and compositional computation [20]. Power systems, as in short-circuit current calculation, impose exactly such stringent requirements on precise multi-step solving. Setting the three reasoning calculation subtasks is therefore particularly important for ascertaining the capability floor of models in engineering practice. The seven subtasks are not isolated but form a capability chain from basic knowledge to engineering application: knowledge question answering asks whether a model knows and understands, whereas reasoning calculation asks whether it can analyze and solve. Through this

structured design, PowerGLUE can identify differences in model performance across capability dimensions beyond the overall evaluation.

Dataset construction and quality control

To ensure that PowerGLUE evaluates the capabilities of LLMs in professional power scenarios as objectively as possible, we built the dataset and the evaluation method around a pipeline of data sourcing, sample screening, categorical annotation, supplementary generation, expert verification, statistical analysis, and automatic scoring (Fig. 1). Unlike general open-domain evaluation, items in power scenarios must be accurate in content, conform to industry conventions in wording, have discriminating option settings, and be traceable to verifiable answer grounds. The construction therefore emphasizes the authority of data sources, the representativeness of the task distribution, and the controllability of item quality.

PowerGLUE draws on four classes of professional material: standards and regulations (industry standards such as GB/T and DL/T, covering terminology definitions, technical clauses, parameter requirements, and equipment specifications); professional examination materials (questions from the registered electrical engineer examination, which introduce objective items with strong discriminating power); textbooks and literature, which supplement knowledge of fundamental principles, equipment mechanisms, and system analysis; and engineering cases, which support application-oriented tasks such as fault analysis, engineering calculation, and planning and design. In the current version, 164 standards have been organized and more than 5,000 authentic registered electrical engineer examination questions have been collected, providing a rich base corpus for item construction. Standards and examination questions guarantee the professionalism, normativity, and objective decidability of items, while textbooks and engineering cases help cover the principle explanations, fault logic, and engineering application problems that pure normative clauses can hardly reach.

After raw data collection, materials from different sources are first unified through format normalization, duplicate removal, splitting of question stems and options, and standardization of answer fields. For objective items collected directly, those with complete stems, unambiguous options, verifiable answers, and professionally compliant wording are retained preferentially; for non-question text from textbooks, standards, and cases, knowledge fragments that can be converted into question-answering or reasoning calculation tasks are extracted for subsequent sample construction. Screening removes four classes of problems: (1) duplicate and near-duplicate samples, whose stems are highly similar in meaning and differ only in numbers or minor wording; (2) incomplete samples, with missing options, ambiguous answers, or insufficient expression of intent; (3) samples with disputed answers, unclear standard grounds, or additional premises required for answering; and (4) open-ended questions that cannot be converted into deterministic options. Through this processing, the raw data are condensed into a candidate sample set better suited to standardized objective evaluation.

After cleaning and preliminary screening, candidate samples are classified and annotated according to the task system described above. Each item carries two types of labels: a business-sector label mapping it to one of the six core sectors, and a task label assigning it first to knowledge question answering or reasoning calculation and then to one of the seven subtasks; this three-level structure ensures that results can be analyzed at different granularities. Given the large scale of the candidate set and the semantic complexity of the items, annotation follows a layered workflow: a large model performs the initial classification, a second model checks the consistency of the classification, and only samples whose two-round labels agree enter the candidate pool; samples with inconsistent labels, samples involving multiple knowledge points, and samples with blurred task boundaries are referred to the expert panel for the final ruling.

In manual review, each business sector is covered by experts with front-line power-industry experience: each sector's expert panel consists of three experts from power supply bureaus, and the final classification of disputed samples is decided by vote, with samples lacking clear consensus or inherently cross-sector with an undetermined primary label excluded from the formal evaluation set. In complex, high-risk engineering fields, relying solely on automated model

processing tends to introduce hard-to-detect systematic bias, and a strict human-experts-in-the-loop mechanism is required to align the underlying professional logic precisely [21]; model-assisted classification serves only to raise annotation efficiency and does not replace human judgment. The prompt template used for classification is given in Supplementary Note S1.

Because directly collected items are unevenly distributed across sectors and subtasks, and samples are sparse in some long-tail scenarios such as dispatching and distribution, we introduce a supplementary generation mechanism beyond the original item sources to improve the coverage of the dataset. To avoid the hallucination contamination inherent in unconstrained free generation [22], this mechanism does not expand the item bank by free generation; instead, using retrieval-augmented generation [23], it follows a pipeline of retrieval constraint, template generation, and manual audit, controlling deviations in knowledge grounds, item structure, and answer uniqueness. In total, 471 supplementary items were generated, accounting for 13.07% of the formal item bank (the generation and auditing procedures are described in Methods).

To guarantee data quality, items undergo manual verification before entering the formal evaluation set. Verification focuses on four aspects: whether the wording of items is accurate, clear, and consistent with the power professional context; whether the sector and task labels are reasonable; whether the options are discriminating, without obvious hints or ambiguity; and whether the standard answers and their grounds are clear and verifiable. For directly collected items, manual verification mainly confirms the consistency of labels and answers; for model-generated items, knowledge provenance, stem wording, and distractor plausibility are additionally checked. Samples whose grounds cannot be clearly located, whose answers are disputed, or whose wording is imprecise are excluded from the formal test set.

Following the pipeline above, PowerGLUE forms an evaluation dataset covering the six business sectors. In the current version, the total numbers of items in generation, transmission, transformation, distribution, consumption, and dispatching are 616, 656, 750, 572, 652, and 357, respectively, of which the knowledge question answering items number 495, 477, 492, 403, 499, and 236, and the reasoning calculation items number 121, 179, 258, 169, 153, and 121 (Table 3). Overall, PowerGLUE achieves basic coverage of the six core business sectors, although item counts still differ across sectors. This imbalance is mainly related to the availability of raw data and the difficulty of standardized item construction in different sectors; we will continue to develop the item bank to maintain inter-sector balance.

Table 3
Distribution of PowerGLUE items across business sectors and task categories.

Sector	Knowledge question answering	Reasoning calculation	Total
Generation	495	121	616
Transmission	477	179	656
Transformation	492	258	750
Distribution	403	169	572
Consumption	499	153	652
Dispatching	236	121	357
Total	2,602	1,001	3,603

Evaluation of LLMs on PowerGLUE

We evaluated seven representative LLMs from the DeepSeek, Kimi, Qwen, and GLM families, covering both instruct and thinking (reasoning) variants (Tables 4 and 5). All models were tested on the same item bank, with the same prompt template and the same answer-extraction rules, under the evaluation protocol described in Methods.

Tables 4 and 5 report the scores of each model on the knowledge question answering and reasoning calculation tasks, respectively. Overall performance differs markedly across models. Qwen3-235B-A22B-Thinking-2507 obtained the highest overall scores, 66.49% on knowledge question answering and 73.63% on reasoning calculation; Kimi-K2-Instruct, DeepSeek-V3.1, DeepSeek-R1-0528, Qwen3-235B-A22B-Instruct-2507, and GLM-4.6 formed a closely spaced middle group, with overall scores on both tracks between approximately 61% and 73%; and Kimi-K2-Thinking obtained markedly lower scores on both task categories, 33.36% and 26.77%, respectively. Because the reported scores come from a single evaluation run, differences of one to three percentage points should be interpreted with caution; even so, the wide overall spread indicates that PowerGLUE separates the evaluated models over a broad range.

Table 4
Accuracy (%) of the evaluated LLMs on the knowledge question answering tasks, by sector.

Model	Generation	Transmission	Transformation	Distribution	Consumption	Dispatching	Overall
DeepSeek-V3.1	50.51	60.17	64.02	68.49	62.53	72.03	61.88
DeepSeek-R1-0528	50.51	57.23	62.80	71.96	61.12	74.58	61.61
Kimi-K2-Instruct	54.14	61.43	68.29	72.95	64.73	75.00	64.99
Kimi-K2-Thinking	29.70	30.61	29.27	37.72	30.46	53.81	33.36
Qwen3-235B-A22B-Instruct-2507	50.91	57.86	66.87	74.69	61.92	76.27	63.30
Qwen3-235B-A22B-Thinking-2507	55.35	60.38	69.92	75.93	67.33	77.12	66.49
GLM-4.6	49.49	59.75	65.04	70.22	55.51	72.88	60.80

Table 5
Accuracy (%) of the evaluated LLMs on the reasoning calculation tasks, by sector.

Model	Generation	Transmission	Transformation	Distribution	Consumption	Dispatching	Overall
DeepSeek-V3.1	80.17	59.22	63.57	88.17	75.51	83.47	73.02
DeepSeek-R1-0528	80.99	57.54	56.98	88.76	74.51	72.73	69.93
Kimi-K2-Instruct	80.17	60.89	60.47	90.53	77.12	79.34	72.83
Kimi-K2-Thinking	41.32	20.11	12.24	44.97	27.88	25.62	26.77
Qwen3-235B-A22B-Instruct-2507	76.03	50.84	50.78	75.74	69.93	71.07	63.44
Qwen3-235B-A22B-Thinking-2507	81.82	63.69	62.40	87.57	78.43	78.51	73.63
GLM-4.6	74.38	52.52	60.07	88.17	67.32	75.21	68.13

A comparison across the two task categories shows that, under the current item composition, most models score higher on reasoning calculation than on knowledge question answering. DeepSeek-V3.1 rises from an overall 61.88% on knowledge question answering to 73.02% on reasoning calculation; GLM-4.6 scores 60.80% and 68.13%; and Qwen3-235B-A22B-Thinking-2507 scores 66.49% and 73.63%. We note that the two tracks are not matched in size, difficulty, or option structure, as they comprise 2,602 knowledge items and 1,001 calculation items drawn from different sources, so this pattern indicates a measurable difference between the tracks under the present benchmark rather than a direct comparison of knowledge and reasoning ability. Even so, the pattern is informative for benchmark design: several models perform well on formula application, parameter conversion, and process deduction, yet show clear shortcomings on knowledge tasks involving standard clauses, professional terminology, and equipment mechanisms. The dual-track design of PowerGLUE thus decomposes model capability and improves the interpretability of the evaluation results.

As shown in Figs. 2 and 3, model performance is uneven across the six business sectors. On knowledge question answering, most models score relatively high in dispatching and distribution and relatively low in generation and transmission. Qwen3-235B-A22B-Thinking-2507 reaches 77.12% in dispatching and 75.93% in distribution, against 55.35% in generation and 60.38% in transmission; DeepSeek-V3.1 scores 72.03% in dispatching and 68.49% in distribution, but only 50.51% in generation. PowerGLUE is thus fairly sensitive at the sector level, reflecting the unevenness of models' coverage of industry knowledge.

On reasoning calculation, sector differences are equally evident. Overall, scores are relatively high in distribution, dispatching, and generation, and relatively low in transmission and transformation. Qwen3-235B-A22B-Thinking-2507 scores 81.82%, 87.57%, and 78.51% in generation, distribution, and dispatching, versus 63.69% and 62.40% in transmission and transformation; DeepSeek-V3.1 reaches 88.17% and 83.47% in distribution and dispatching, but only 59.22% and 63.57% in transmission and transformation. Even within the same task category, model capability is thus affected by the concrete business scenario.

These results are consistent with the design intent of PowerGLUE in two respects. First, the six business sectors are not a mere listing of scenarios: they genuinely distinguish model capability across different links of the power system. Second,

by organizing knowledge question answering and reasoning calculation uniformly within each sector, PowerGLUE can simultaneously expose task differences within a sector and the distributional differences of a task across sectors, and thus provides a finer-grained capability profile.

Beyond overall accuracy, PowerGLUE is also intended to expose common problems across capability dimensions. Combining the observations above, most evaluated models obtain relatively higher scores on calculation tasks and lower scores on tasks involving standard clauses, professional concepts, and equipment principles. Under the present item composition and scoring protocol, this pattern is consistent with the interpretation that the knowledge formed through general pre-training is not yet sufficient to support the highly constrained and strongly normative application needs of the power industry; however, because the two tracks differ in difficulty and composition and because non-conforming outputs are counted as incorrect, the pattern should not be read as a direct causal comparison between knowledge and reasoning ability. PowerGLUE supports this decoupled observation of knowledge-track versus calculation-track performance.

PowerGLUE also separates different reasoning modes. Several thinking-mode models hold potential advantages in complex reasoning tasks, but when their output format is unstable or instruction following is inadequate, their final scores drop markedly because of answer-extraction failures. Kimi-K2-Thinking, for instance, obtained lower exact-match scores than the instruct version of the same family on both task categories. Because invalid or non-conforming outputs were counted as incorrect, this result reflects the combined effects of task performance, instruction following, and answer extractability under the present prompting and parsing protocol, and should not be interpreted as evidence that thinking-mode reasoning is intrinsically less effective. It does indicate that, for professional applications that rely on automated pipelines, reasoning depth and output regularity are both consequential, and PowerGLUE scores partly reflect the latter.

Taken together, the experimental results indicate that PowerGLUE can distinguish models in professional power scenarios in three respects: the overall capability levels of different models, the unevenness of model performance across sectors and task types, and common weaknesses related to knowledge mastery, engineering reasoning, and output regularity. PowerGLUE can therefore serve not only as a testbed for horizontal model comparison but also as a reference point for subsequent domain fine-tuning, data augmentation, and engineering adaptation.

Discussion

Focusing on capability evaluation of LLMs in the electric power industry, this study constructed PowerGLUE and analyzed seven mainstream models through its dual-track task system. The results show that model capabilities in professional power scenarios are uneven across task types, business sectors, and reasoning modes: under the present item composition, most models score higher on reasoning calculation than on knowledge tasks involving professional understanding, standard-clause mastery, and equipment-mechanism cognition (a pattern to be interpreted with caution given differences in item composition and difficulty), and some thinking-mode models lose marks when unstable output formats cause answer-extraction failures. Evaluating LLMs in the electric power industry therefore cannot simply borrow the approach of general language tasks; it should target the industry's key capabilities, including professional knowledge, regulation understanding, engineering reasoning, and output stability. As an initial benchmark oriented to professional power applications, PowerGLUE provides a unified test framework for subsequent model comparison, domain fine-tuning, and industry-adaptation research; continued improvement in data scale, task complexity, evaluation dimensions, and reproducibility remains needed.

Methodologically, the value of PowerGLUE is threefold. First, compared with general-purpose benchmarks, PowerGLUE emphasizes adaptation to the scenarios of the electric power industry and covers the typical knowledge and application problems along the main business chain of power systems. Second, through the combination of knowledge question answering and reasoning calculation, it can to a certain extent separate what a model knows from what a model can do. Third, through multi-level analysis across sectors, tasks, and models, it supports not only horizontal comparison but also targeted diagnosis for subsequent domain fine-tuning, data augmentation, and engineering adaptation. The significance of

a professional benchmark lies not merely in producing an overall score, but in helping researchers understand the capability boundary of models in real industry scenarios.

Nevertheless, the current version of PowerGLUE has limitations. First, although the dataset covers the six core business sectors, the distribution of items across sectors is not yet fully balanced; the number of items in some scenarios, especially dispatching, is relatively small, which may affect the interpretability of the overall results. Second, the current evaluation consists mainly of objective items, which facilitates unified scoring and improves evaluation efficiency, but its portrayal of open-ended analysis, complex decision chains, and multi-step reasoning processes remains limited. Third, the construction process introduces model-assisted classification and generation; although some bias has been controlled through manual verification and standard-answer matching, transparency regarding data provenance, annotation consistency, and the share of generated items still needs to be strengthened to improve the reproducibility and credibility of the benchmark. Fourth, the current scoring is based mainly on answer matching, and finer assessment of the soundness of the reasoning process and the correctness of intermediate steps is not reflected in the scores. Fifth, the reported scores are based on a single evaluation run per model without repeat-run variance estimates or confidence intervals, and outputs that fail to conform to the prescribed answer format are counted as incorrect, so the reported accuracies combine content correctness with format compliance; small differences between models should therefore be interpreted with caution.

Future work can proceed along four directions. First, continue to expand high-quality power-domain data to further balance the distribution across sectors and subtasks. Second, while retaining objective scorability, introduce more complex task forms, such as case analysis, regulation deduction, and multi-step engineering calculation, to deepen the portrayal of real business capability. Third, improve the public documentation of dataset construction and evaluation protocols, including annotation rules and statistics on extraction failures, to strengthen openness and reproducibility. Fourth, in view of the application characteristics of the power industry, explore incorporating result correctness, output regularity, instruction following, and process interpretability into a unified evaluation framework, extending PowerGLUE from result evaluation toward an industry benchmark that weighs capability and behavior equally.

Methods

Supplementary item generation and auditing

We first build a power knowledge base from standards and regulations, textbooks and literature, and engineering cases. The knowledge base takes traceable knowledge fragments as its basic units, mainly including standard clauses, equipment-principle descriptions, calculation rules, and engineering case descriptions. The previously collected registered electrical engineer examination questions are classified and used as the base draft of the dataset, serving to identify sample gaps and to inform the design of generation templates. For sectors and subtasks with an insufficient number of samples, candidate objective items are then generated in a retrieval-augmented manner: the knowledge fragments most relevant to the target task are retrieved from the knowledge base and combined with a unified prompt template. The template explicitly constrains the item's sector, task category, answer uniqueness, number of options, and distractor setting, so as to reduce errors caused by unsupported generation and open-ended phrasing. The prompt template used for generation is given in Supplementary Note S2.

In auditing generated items, we verify that the question stem is consistent with the retrieved knowledge grounds, that the option setting is complete and discriminating, that the standard answer is unique and traceable, and that the item can be stably assigned to the established sector and task categories. Candidates with unclear knowledge provenance, disputed answers, wording inconsistent with the power professional context, or high similarity to existing samples are all excluded from the formal evaluation set. The prompt template used for auditing is given in Supplementary Note S3. Through this process, generated items serve only as a supplement to the original sources and do not replace authoritative data such as standards and examination questions. In total, 471 supplementary items were generated, accounting for 13.07% of the

formal item bank. This share is deliberately kept low: it improves the coverage of long-tail tasks while preventing the formal evaluation set from depending excessively on model-generated data. It should further be noted that although PowerGLUE uses large models to generate some candidate items at the construction stage, it does not use large models as scoring judges at the evaluation stage; scoring is performed automatically by matching against standard answers. The potential bias of generated items is therefore controlled through knowledge-ground tracing, manual review, and objective answer determination, rather than turning into an open-ended problem of models judging models.

Evaluation protocol and scoring

To allow different models to be compared under unified conditions, PowerGLUE organizes all evaluation tasks in the form of objective items and uses automated scripts for answer extraction and scoring. In the evaluation, we adopt zero-shot prompting [24]: each question is input to the model, and a unified format constraint requires the model to output its final option. The prompt requires the model to give its answer option in a fixed format after analysis; the scoring script then extracts the final answer from the model output using regular expressions and compares it with the standard answer. The total score of a model is computed as the proportion of correctly answered items among all items. The prompt template used in the current version is given in Supplementary Note S4.

The use of objective items and rule-based matching is intended to reduce the inconsistency of evaluation criteria inherent in open-ended answers and to improve the operability and repeatability of comparisons across models. It should be pointed out that this approach cannot completely eliminate bias and may still be affected by unstable output formats, models not answering as required, and parsing failures of multiple answers. If a model fails to output in the prescribed format, or gives no answer within the token limit, the output is treated as invalid and counted as incorrect in the reported accuracies. This strict rule reflects deployment-oriented requirements on instruction following and answer conciseness, and implies that the reported scores combine content correctness with format compliance.

Experimental settings

To verify the effectiveness of PowerGLUE for evaluation in professional power scenarios, we selected several representative LLMs for experimental testing. Because models differ in parameter scale, training corpus, reasoning mode, and output habits, the experiments adopt unified input and scoring procedures wherever possible to keep the results comparable. All models were tested on the same item bank, with the same prompt template and the same answer-extraction rules; the final options in model outputs were extracted by automated scripts and compared with the standard answers, and the proportion of correctly answered items was taken as the model's accuracy on the corresponding task.

In implementation, the models were called through their APIs. Temperature was uniformly set to 0.7, the maximum number of tokens to 8192, and the thinking budget to 4096; for models supporting reasoning or thinking modes, the thinking mode was used. Each question was input to the model with the unified prompt template, and the model was required to output its final option in a fixed format after analysis, reducing the interference of free-form expression with automated scoring. The scoring script uses regular-expression matching to extract final answers and computes model performance on the knowledge question answering tasks, the reasoning calculation tasks, and each business sector separately. The current experiments are based mainly on the results of a single complete run: given the large number of items, most of them objective, a single run under the unified prompt template and output-format constraints can reasonably reflect a model's overall capability level on the benchmark; repeated testing and variance analysis will be introduced in future work to assess result stability. For samples in which the model does not output a final answer in the prescribed format and the automated script cannot stably extract a valid option, the output is treated as invalid, counted as incorrect in the reported accuracies, and discussed in the analysis together with the model's output regularity.

During the preparation of this manuscript, the authors used a large language model to assist with English translation and language editing; the manuscript was originally drafted in Chinese. All scientific content, data, and conclusions were authored and verified by the authors, who take full responsibility for the work.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study involved power-industry domain experts in the annotation and verification of benchmark items; informed consent was obtained from all participating experts prior to their participation. No personal data, biological samples, or experimental interventions were involved.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

C.Z., C. P., S. L., and Z. C. conceived and designed the study. C. Z., Z. P. and Z. C. helped build the dataset and perform the experiments. C.Z. wrote the manuscript.

Data availability

The PowerGLUE dataset is available upon reasonable request. A public online evaluation platform for the benchmark is currently under development and will be made available in the future.

Code availability

The prompt templates used for item classification, supplementary item generation, item auditing, and evaluation are provided in Supplementary Notes S1-S4 of the Supplementary Information.

References

1. Shu, Y., Chen, G., He, J. & Zhang, F. Building a new electric power system based on new energy sources. *Chin. J. Eng. Sci.* **23**, 61 (2021).
2. Bernadić, A., Kujundžić, G. & Primorac, I. Large language models in power systems: enhancing control and decision-making. *Int. J. Innovative Solutions Eng.* **1**, 10–17 (2025).
3. Mirshekali, H., Shadi, M. R., Ghanadi Ladani, F. & Shaker, H. R. A review of large language models for energy systems: applications, challenges, and future prospects. *IEEE Access.* **13**, 163162–163188 (2025).
4. EPRI. *Benchmarking Large Language Models for the Electric Power Sector* (Report (Electric Power Research Institute, 2025).
5. Amjad, F., Korōtko, T. & Rosin, A. Review of LLMs applications in electrical power and energy systems. *IEEE Access.* **13**, 150951–150969 (2025).

6. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2018).
7. Wang, A. et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Adv. Neural. Inf. Process. Syst.* **32** (2019).
8. Hendrycks, D. et al. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).
9. Huang, Y. et al. C-Eval: a multi-level multi-disciplinary Chinese evaluation suite for foundation models. *Adv. Neural. Inf. Process. Syst.* **36**, 62991–63010 (2023).
10. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
11. Guo, X. et al. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 6258–6292 (2025).
12. Wu, J. et al. Arch-Eval benchmark for assessing chinese architectural domain knowledge in large language models. *Sci. Rep.* **15**, 13485 (2025).
13. Chang, Y. et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
14. Zhou, X. et al. ElecBench: a power grid dispatch evaluation benchmark for large language models. In Proceedings of the IEEE Power and Energy Society General Meeting (PESGM), 1–5 (2025).
15. Zhang, S. et al. ElecBench: a large language model benchmark in electric power domain. *Eng. Appl. Artif. Intell.* **162**, 112310 (2025).
16. Wang, Y. et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. *Adv. Neural. Inf. Process. Syst.* **37**, 95266–95290 (2024).
17. Vieira et al. InsPLAD: a dataset and benchmark for power line asset inspection in UAV images. *Int. J. Remote Sens.* **44**, 7294–7320 (2023).
18. Tao, X. et al. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Trans. Syst. Man. Cybernetics: Syst.* **50**, 1486–1498 (2020).
19. Bommasani, R., Liang, P. & Lee, T. Holistic evaluation of language models. *Ann. N. Y. Acad. Sci.* **1525**, 140–146 (2023).
20. Dziri, N. et al. Faith and fate: limits of transformers on compositionality. *Adv. Neural. Inf. Process. Syst.* **36**, 70293–70332 (2023).
21. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural. Inf. Process. Syst.* **35**, 27730–27744 (2022).
22. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surveys.* **55**, 1–38 (2023).
23. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural. Inf. Process. Syst.* **33**, 9459–9474 (2020).
24. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural. Inf. Process. Syst.* **35**, 22199–22213 (2022).

Figures

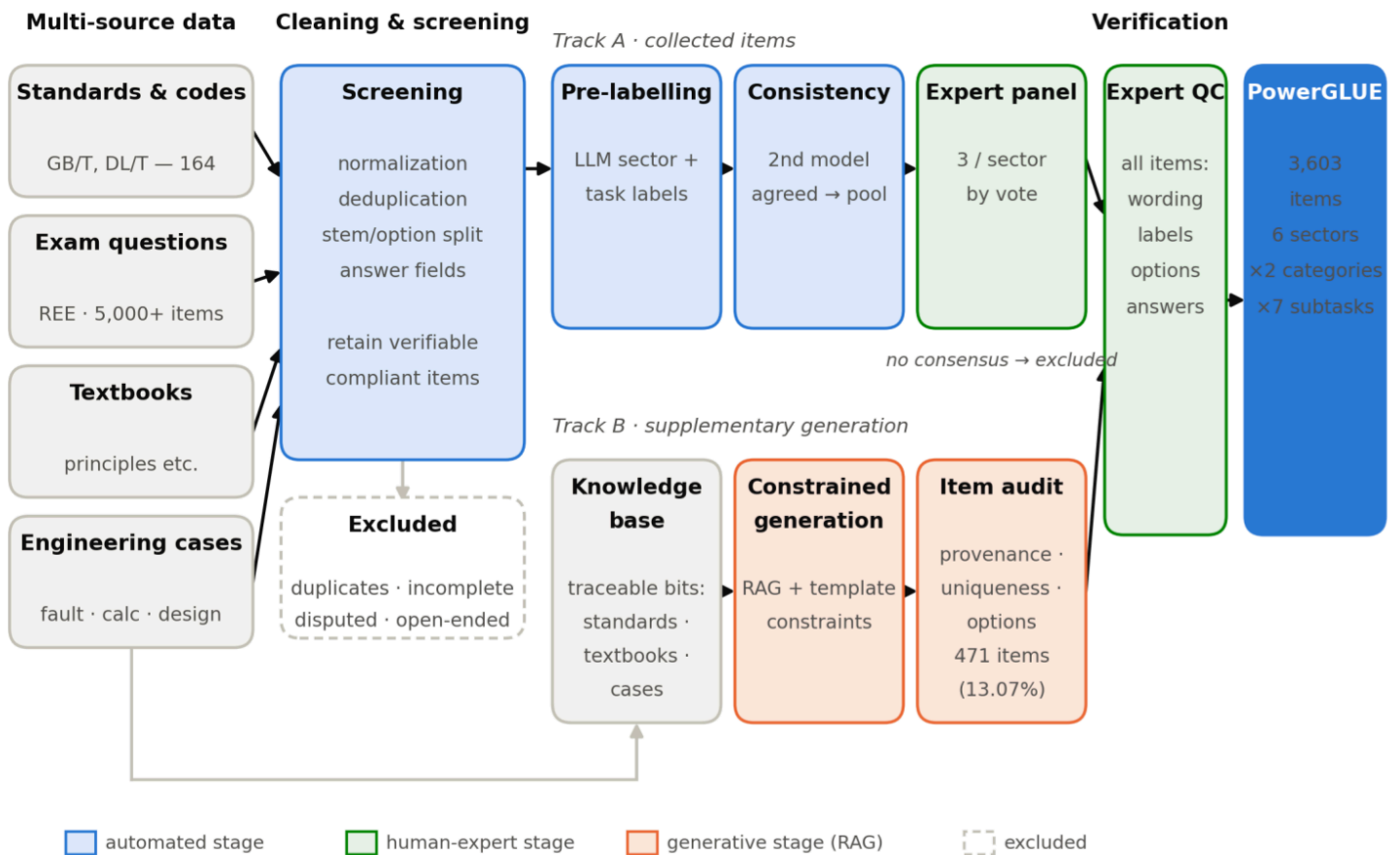


Figure 1

Construction pipeline of the PowerGLUE benchmark. Multi-source professional materials (164 standards and regulations, over 5,000 registered electrical engineer examination questions, textbooks and literature, and engineering cases) are cleaned and screened. Directly collected items are labelled through two-model consistency classification with expert-panel adjudication (Track A); a retrieval-constrained generation stage over a traceable power knowledge base supplements sparse sectors with 471 audited items, 13.07% of the bank (Track B). All items pass expert verification before entering the benchmark, which contains 3,603 items spanning six sectors, two task categories, and seven subtasks.

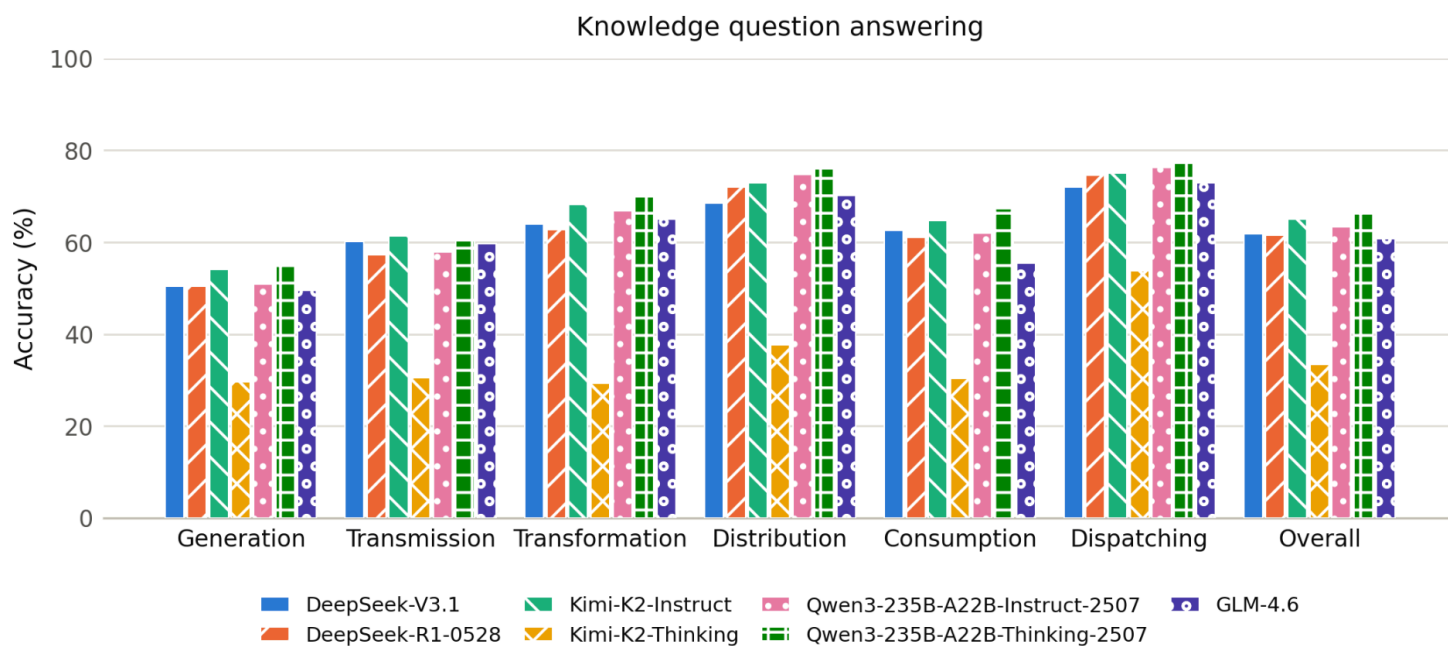


Figure 2

Domain-level accuracy of the seven evaluated LLMs on the knowledge question answering tasks.

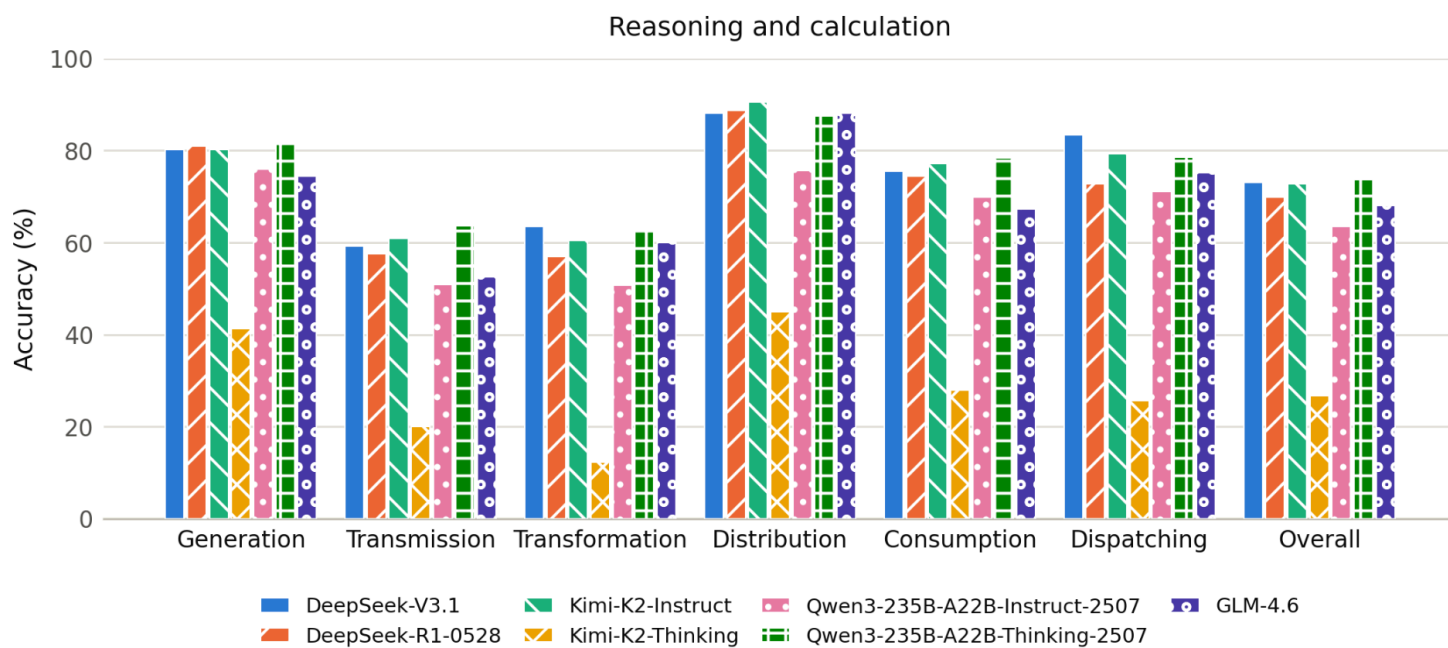


Figure 3

Domain-level accuracy of the seven evaluated LLMs on the reasoning calculation tasks.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PowerGLUEScientificReportsSI0909.docx](#)