Trustworthy Large Language Models in Organisations: A Competence-Based Alignment Framework

Daniel Pascal Hefti* and Tianxiang Lu*
*IU International University of Applied Science, Erfurt, Germany
Email: daniel.hefti@gmail.com, tianxiang.lu@iu.org

***Abstract*—Large language models (LLMs) increasingly influence strategic decision-making in organisations, yet their alignment with organisational culture, values, and goals (OCVG) remains underexplored. Existing bias mitigation techniques primarily address societal categories—such as gender, race, and religion—offering limited relevance for small and medium-sized enterprises (SMEs) seeking to embed their unique ethos into AI systems. Moreover, SMEs often lack the resources to implement complex alignment strategies. This study proposes a novel, human-centred approach: adapting methods from competence management — traditionally used to align human behaviour with organisational expectations — to guide and evaluate LLM outputs. Results show statistically significant improvements in alignment scores for treatment models across different datasets and languages (Mann–Whitney U, $p < .001$; Rank Biserial Correlation $r_{rb} = .36$). Human validation of 8.73% of outputs confirmed strong agreement with automated evaluations (Pearson $r = .71$, $p < .001$), underscoring the reliability of our rubric-driven assessment. These findings demonstrate that competence-based instructions offer a scalable, interpretable, and ethically grounded method for aligning LLMs with organisational values.**

***Index Terms*—Large language models, Bias mitigation, Bias detection, AI governance, Organisational culture, Competence models**

I. Introduction

Large language models (LLMs) have rapidly transitioned from research prototypes to core enterprise tools. Recent surveys indicate that 72% of organisations report using AI, with 65% regularly leveraging generative AI in business functions, reflecting rapid adoption and investment growth [1]. Despite widespread use, only about 31% of organisations have comprehensive AI policies, revealing significant governance gaps [2]. SMEs in particular often lack the expertise or resources to manage AI risks effectively.

LLMs introduce unique operational and strategic risks. They generate plausible but sometimes incorrect outputs, may reproduce undesirable patterns from training data, and can produce unpredictable behaviour that affects decision quality, compliance, and strategic alignment [3], [4]. Without effective alignment, LLM outputs may inadvertently undermine organisational priorities or conflict with cultural norms.

In organisational science, culture encompasses shared values, beliefs, and assumptions that shape behaviour and decision-making within a firm [5], [6]. Values and goals define normative expectations about priorities and desired outcomes, and their alignment with individual actions has been shown to influence organisational performance [7]. In this study, we refer to *organisational culture, values, and goals* (OCVG) as the constellation of strategic priorities, behavioural norms, and normative expectations that guide action within an organisation.

Existing research on LLM alignment primarily focuses on mitigating bias along broad societal axes, such as gender, race, or ethnicity [3]. While essential for ethical compliance, these approaches do not address alignment with an individual organisation's culture, values, and goals. Small and Medium Enterprises (SMEs) require practical, interpretable alignment methods that do not demand advanced AI expertise.

We propose a novel approach: adapting **competence management frameworks** to guide and evaluate LLM outputs. Competence models are traditionally used to translate desired behaviours into actionable standards for human employees [8]. They operationalise abstract values as behavioural anchors, providing structured guidance for normative alignment. By grounding alignment criteria in organisational priorities, we explore whether strategies designed for human alignment can improve the alignment of LLM behaviour with OCVG.

Contributions

- We introduce **competence-based textual instructions** as a practical, non-technical **alignment strategy** for LLMs.
- We **validate rubric-driven, LLM-as-a-Judge evaluation** as a scalable method for assessing alignment with OCVG.
- We ground automated evaluations through **human validation** of 8.73% of model responses.
- We **provide experimental evidence** that competence-based instructions improve alignment with OCVG across tasks, languages, and datasets.

II. Related Work

A. Organisational Values and Competence-Based Alignment

Organisational culture, values, and goals (OCVG) define the shared norms and strategic priorities that guide behaviour within a firm [6], [9]. Established frameworks such as the McKinsey 7-S model and the Competing Values Framework

link cultural attributes to organisational performance [10]–[12], but operate at an aggregate level and do not specify how desired behaviour can be operationalised for individual actors.

Competence models address this limitation by translating abstract values into behavioural anchors—observable patterns that define how values should manifest in practice [13], [14]. These models have been shown to improve organisational performance, commitment, innovation, and the quality of decisions by aligning individual behaviour with strategic priorities [15]–[18], as well as providing ethical reference points [19]. Crucially, competence models regulate externally observable behaviour rather than internal cognition, making them conceptually transferable to black-box systems such as LLMs, where internal reasoning processes are inaccessible.

B. LLM Bias Mitigation

LLMs often reproduce societal stereotypes embedded in their training data and algorithms. Among existing mitigation strategies, only a small subset is feasible for SMEs without AI expertise. Generated-text-based metrics and textual instructions [20] offer practical options because they require minimal expertise and work with black-box models [3]. Prior work showed that diversity-promoting textual instructions didn't reduce bias [21], but did not address alignment with OCVG [5]. While bias mitigation addresses external social harms, organisations face an additional internal governance problem: ensuring that LLM behaviour remains normatively consistent with their own strategic and ethical commitments.

Complementing bias mitigation, *guardrails* provide explicit constraints—implemented externally or within prompts—to enforce safe and norm-consistent model behaviour [22], [23]. While bias mitigation primarily addresses external social harms, guardrails focus on reducing risky or undesirable outputs through rule-based constraints. In contrast, our OCVG-based approach explicitly aligns LLM behaviour with an organisation's specific values, priorities, and ethical commitments, ensuring normative consistency that goes beyond general safety or fairness measures.

C. Evaluating LLM Responses

Contemporary evaluation approaches increasingly rely on LLM-as-a-Judge systems [24], in which one model evaluates another. Within this paradigm, reference-free frameworks such as retrieval augmented generation assessment (RAGAS) reduce resource demands by assessing faithfulness and relevance without ground-truth annotations, yet remain task-agnostic and cannot verify adherence to organisational norms [25]. In contrast, rubric-based evaluation represents a conceptual shift: by introducing structured scoring criteria, it enables fine-grained and interpretable assessments. Recent LLM-as-a-Judge systems, including Prometheus and Hercule [26]–[28], operationalise this approach and approximate human judgment under constrained, rubric-defined conditions. They are scalable and support custom rubrics for competence-based evaluation, but do not prescribe how evaluation criteria should be derived or aligned with OCVG. Moreover, evaluator bias remains a persistent challenge, underscoring the need for mitigation strategies and periodic human validation of results [29], [30].

D. Positioning of This Work

The literature identifies three largely separate strands: (i) bias mitigation addressing societal harms, (ii) competence models translating organisational values into human behaviour, and (iii) evaluation frameworks measuring task performance via automated or rubric-based methods. While complementary, these strands remain disconnected—methods either guide behaviour without organisational grounding or assess outputs without normative alignment. Table I situates our approach within this landscape.

TABLE I: Comparison of LLM Alignment Approaches

Approach	**Target**	**Mechanism**	**Interp.**	**Eval.**
Bias mitigation [3], [4], [20], [21]	Societal	Internal and Prompt	Low	Generic
Guardrails [22], [23]	Task / Safety	External (Rule-based)	Medium	Rule-based
RAGAS [25]	Task	External	High	Task-based
LLM-as-a-Judge [24], [26]	Task	External	Medium	Task-based
This study	**OCVG**	**External (Prompt)**	**High**	**Normative**

III. Methodology

A. Design and Hypotheses

This study employs a randomized controlled trial (RCT) following a pretest–posttest design with intervention withdrawal [31]. In this RCT, LLMs are our participants, and the overall objective of the experiment is to verify if the treatment improves alignment with OCVG. The experiment proceeds through three sequential stages—baseline, intervention, and withdrawal— corresponding to an $A_1XA_2XA_3$ structure. Competence-based textual instructions are introduced exclusively during the intervention stage and removed prior to the withdrawal stage. This design enables within-model comparison across stages and strengthens causal inference through reversibility testing. Because models are stateless across runs and receive no persistent memory, withdrawal of instructions restores baseline conditions, satisfying the reversibility assumption.

The experiment is implemented across two phases to balance reproducibility and contextual validity. The first (reproducible) phase uses a public English dataset, while the second (contextual) phase employs a proprietary German dataset from a participating SME. During the intervention stage, treatment models receive structured textual instructions derived from the SME's competence model. These instructions operationalise organisational values as behavioural anchors, defined as observable, task-specific criteria that specify how abstract values should manifest in practice [13].

The alternative hypothesis (H_1) posits that textual instructions communicating desired competencies significantly enhance alignment with OCVG, whereas the null hypothesis (H_0) assumes no meaningful difference between treatment and control groups.

B. Sampling Procedure

In order to conduct the experiment, we must apply a sampling procedure three times. First, we need to select the models that will participate in the experiment. Then, we need to select the models that will receive the treatment, and finally, the responses that will undergo human evaluation. Stratification is used to ensure that models and responses are distributed equally. The sampling procedure was implemented in Python using the pandas GroupBy.sample() method [32]. From each stratum, a simple random sample of N units was selected. This constitutes a deliberate disproportionate allocation, which is designed to ensure equal representation across strata. This approach considers Cochran's observation that equal sample sizes can yield greater precision than proportional allocation [33].

C. Models, Tasks, and Datasets

The study uses stratified random sampling to select ten out of fifty open-source LLM models for the experiment, and to select five of these for the treatment group. Architectural diversity is ensured by sampling across five parameter-size strata ranging from 135 million to 671 billion parameters. All models are treated as black boxes. None are fine-tuned, modified, or granted access to internal state. The unit of analysis is model output rather than model count, yielding several thousand evaluated responses across phases and conditions.

A single task, sales analytics for a head of sales, was selected in consultation with a participating SME for its strategic relevance and capacity to elicit competence-related behaviours. It requires models to analyse aggregated transactional data and generate actionable recommendations on regional performance, customer segmentation, and retention strategies. A single task was deliberately chosen to control task-induced variance and isolate the effect of competence-based instructions.

Two datasets support the dual-phase design. The reproducible phase uses the public Online Retail II dataset [34], while the contextual phase uses a proprietary German dataset supplied by the participating SME. Both datasets are pre-aggregated using a token-efficient strategy to ensure consistent input structure across models with different context lengths.

All data and materials used in the public phase of this study are openly available on Github[1].

D. Controlled Execution Environment

All experiments were executed in a controlled, containerised environment to prevent version drift, external interference, and uncontrolled stochastic variation. All candidate LLMs were treated as black-box systems and executed via Ollama, ensuring deterministic control of model versions, runtime dependencies, and decoding parameters [35]. Except for one 671B model executed via Ollama Cloud [36] due to hardware feasibility, all models ran under equivalent conditions without systematic differences between treatment and control groups.

To isolate the causal effect of competence-based textual instructions, decoding parameters were fixed across all models, datasets, phases, and stages. This choice prioritises internal validity over model-specific optimisation, a trade-off which is discussed in section VI. Evaluation models were deployed independently using vLLM docker containers [37], applying the configuration recommended by their authors [28], and did not participate in the experiment to avoid self-preference bias [30].

E. Prompt Structure

All prompts followed a fixed modular structure comprising three components.

`{instruction}{task}{context}`

The treatment-group models received explicit instructions during the intervention stage, while the tasks and context were identical for all models across all stages. In the reproducible phase, instructions and tasks were provided in English, and in the contextual phase, tasks and instructions were communicated in German. The instruction position was held constant across all conditions to avoid positional bias [29].

F. Competence-Based Instructions

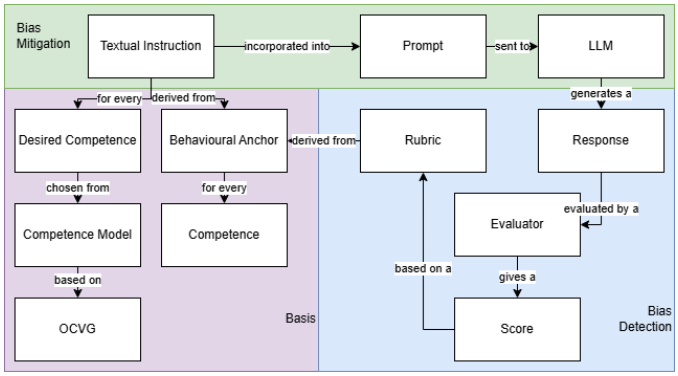

Fig. 1: Conceptual Framework Diagram.

The intervention operationalises organisational values through structured textual instructions derived from the SME's competence model (see Fig. 1). The competence model defines a set of values and corresponding behavioural anchors that describe observable behaviours expected in practice. For the intervention, only those anchors relevant to the sales-analytics task—such as collaboration, transparency, and strategic foresight—were translated into concise instructions and embedded in the treatment prompts.

These anchors also inform the construction of the evaluation criteria used to assess the models' outputs. This approach

[1]https://github.com/Prof-it/ai-alignment

creates a coherent, task-specific, and OCVG grounded instruction set that converts abstract cultural values into actionable behavioural expectations.

While derived from a single SME, the competence categories used (e.g., collaboration, transparency, strategic foresight) reflect widely recognised organisational values. To illustrate the relationship between value, behavioural anchor, instruction, and rubric criterion, an example is provided below.

- **Value**: Collaborative.
- **Behavioural anchor**: I seek solutions that offer added value for everyone.
- **Instruction**: Be collaborative: Seek solutions that offer added value for all stakeholders.
- **Rubric criterion**: To what extent do the recommended measures reflect collaborative thinking and aim to create added value for all stakeholders (e.g., customers, employees, suppliers)?

Table II provides an illustrative example of the rubric scoring scheme. It demonstrates how individual scoring criteria are applied based on the defined evaluation dimensions. The complete rubric is available in our GitHub repository [38].

TABLE II: Rubric for Collaboration Criteria

Grade	Description
1 (Low)	Measures are one-sided and fail to consider the interests or benefits of other stakeholders. They do not foster trust or collaboration.
2	Measures show limited consideration for mutual benefit. While they may offer value to some, they overlook key stakeholder relationships or trust-building aspects.
3	Measures generally aim to benefit multiple stakeholders and show some collaborative intent, but may lack depth in addressing relational dynamics or trust.
4	Measures consistently reflect collaborative thinking and strive to create added value for all relevant parties. They support trust and relationship-building, though minor gaps may remain.
5 (High)	Measures exemplify deep collaborative thinking, clearly aiming to generate mutual value and strengthen trust-based relationships across all stakeholder groups.

G. Evaluation and Analysis

The evaluation strategy combines automated rubric-based assessment with targeted human validation. Model outputs are scored on a five-point Likert scale using structured rubrics derived from the SME's competence model, comprising seven behavioural criteria. Aggregating this seven criteria into a compound alignment score yields total scores between 7 and 35, spanning a total range of 28 points. To justify this aggregation, internal consistency is assessed using Cronbach's α, with values $\geq$.80 indicating acceptable reliability [39]. Two open-source LLM-as-a-Judge systems—M-Prometheus and Hercule—perform automated scoring under rubric-defined constraints. A stratified random sample of 8.73% of all responses was evaluated by humans using the same rubric to benchmark automated scoring [33].

Statistical analysis proceeds in four steps. Descriptive statistics summarise central tendency and dispersion. The Shapiro–Wilk test examines distributional assumptions. Depending on normality, either parametric or non-parametric tests are applied, with Mann–Whitney U tests used for group comparisons under non-normality [40] [41]. Effect sizes are reported using rank-biserial correlation and Cohen's d [42] [43]. Inter-rater agreement between human and automated evaluations is assessed using Pearson correlation, Spearman correlation, and Cohen's κ [44] [45] [46].

IV. Results

The experiment was structured into two phases, to balance reproducibility with contextual relevance. The public phase used an English dataset [34] to ensure replicability, while the contextual phase employed a proprietary German dataset from an SME to test real-world applicability. Each phase comprised two runs:

- **Run 1**: Evaluation *without reference answers* provided to LLM-as-a-Judge systems.
- **Run 2**: Evaluation *with reference answers* provided to LLM-as-a-Judge systems.

These runs allow comparison between reference-free and reference-guided scoring. All reported findings are based on the compound score, which aggregates the seven rubric criteria into a single measure of overall alignment. This aggregation is statistically valid: Cronbach's α exceeded 0.92 for each phase and run (Table III), confirming high internal consistency and justifying the composite metric. Both mean and median are reported to provide a complete view of central tendency, particularly under non-normal distributions. The '-S' suffix indicates that the grades provided by the Hercule evaluator model for the reproducible phase were excluded because they were not statistically significant ($p > 0.1$).

A. Impact of Competence-Based Instructions on Alignment

Competence-based instructions consistently increased alignment scores across all phases and runs (Table III). Treatment group models achieved higher mean and median compound scores than control group models. Across all phases and runs, the median score increased from 19 to 23, and the mean score from 19.2 to 24.74. Effect sizes indicate that the observed gains are not only statistically significant but likely meaningful in operational decision-making contexts for SMEs, where even moderate alignment improvements can influence strategic outcomes.

Fig. 2 illustrate the treatment effect: alignment scores for treatment models rose during the intervention stage and returned to baseline once the instructions were withdrawn. Control models showed no such shift, indicating a genuine treatment effect rather than random variation. Variability, measured by inter-quartile range and standard deviation, did not exhibit an observable pattern, indicating that competence-based prompts shift central tendency without affecting dispersion. Notably, including reference answers led to lower mean and median scores. This effect may reflect anchoring, where scores given by evaluators are based on similarity to the provided reference answer rather than adherence to rubrics.

TABLE III: Comparative Analysis of the Treatment and Control Groups

	α	Mean		Median		IQR	
		Control	Treat.	Control	Treat.	Control	Treat.
RPR1	0.95	23.55	**25.48**	**29.50**	29.00	11.10	**8.85**
RPR1-S	0.93	25.31	**29.78**	29.00	**31.00**	**4.08**	4.10
RPR2	0.92	18.03	**22.09**	19.00	**21.00**	**7.07**	7.25
RPR2-S	0.94	19.98	**25.40**	21.50	**22.50**	3.70	**2.95**
CPR1	0.94	18.68	**23.46**	19.00	**22.00**	7.21	**5.25**
CPR2	0.94	12.77	**19.71**	13.00	**17.00**	**4.75**	6.00
APAR-S	0.95	17.96	**23.64**	18.00	**22.00**	8.01	**7.25**
APAR	0.95	19.20	**24.74**	19.00	**23.00**	**9.33**	9.55

α = Cronbach's α, Treat. = Treatment, RP = reproducible phase, CP = contextual phase, R = run, APAR = all phases and runs. S = statistically significant results only. The results of the winner groups are in bold.

Consequently, responses that diverge from the reference answer, yet still demonstrate robust alignment with the desired values and behaviours, may be systematically undervalued.

The Shapiro–Wilk test confirmed non-normality across all criteria and the compound score ($p < .001$), demanding nonparametric testing. Mann–Whitney U tests revealed statistically significant differences ($p < .01$) between treatment and control groups in all runs except one: the public phase without reference answers when Hercule was used as evaluator model ($p > .10$). Excluding Hercule from this phase produced significant results ($p < .01$), with a rank biserial correlation ranging from 0.29 to 0.43 and Cohen's d ranging from 0.534 to 0.955, indicating moderate to large treatment effects [47]. Both rank-biserial correlation and Cohen's d are reported to provide complementary perspectives on effect magnitude under nonparametric and parametric assumptions.

Given the exploratory yet controlled design and the small number of comparisons, no formal correction for multiple testing was applied; the reported p-values should be interpreted in context of the overall experimental framework.

B. Validity of Rubric-Based Evaluation

Rubric-driven assessment demonstrated strong internal consistency and high agreement with human ratings, confirming its methodological robustness (Table IV). M-Prometheus achieved substantial alignment with human evaluators (Pearson $r = .71$, $p < .001$), while Hercule delivered moderate agreement in the contextual phase (Pearson $r = .58$, $p < .01$). An additional analysis of pairwise comparisons underscores this pattern, with Cohen's κ reaching .732 for M-Prometheus versus .680 for Hercule.

The run with reference answers in the reproducible phase showed marginally higher agreement with human judgements (0.34%) than the run without reference answers (Table V). Whereas, in the contextual phase, reference answers had the opposite effect: agreement was 7.08% higher when reference answers were not provided than when they were. These differences are consistent across evaluators, and indicate that reference-free scoring can perform as well or better than reference-guided scoring in domain-specific contexts.

TABLE IV: Correlation Between Human And LLM Evaluation

	Prometheus		Hercule	
	Pearson	Spearman	Pearson	Spearman
RPR1	0.772***	0.682***	0.036	0.117
RPR2	0.815***	0.733***	0.228	0.274
CPR1	0.771***	0.781***	0.627***	0.635***
CPR2	0.726***	0.637***	0.578**	0.552**
APAR	0.707***	0.695***	0.370***	0.424***

*Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.*

TABLE V: Pairwise Agreement Between Human- and LLM-Evaluation

	Prometheus		Hercule	
	Agreement	Cohen's κ	Agreement	Cohen's κ
RPR1	72.33%	0.707	46.67%	0.438
RPR2	72.67%	0.711	57.00%	0.549
CPR1	74.46%	0.732	69.54%	0.680
CPR2	67.38%	0.658	62.46%	0.606

C. Qualitative Comparison: Impact of Competence-Based Instructions

To illustrate the practical effect of competence-based instructions, we present a representative sales-analytics scenario from our experiment. A head of sales must analyse performance changes in sales regions ranging from top performers to regions with high customer churn and underperforming staff, and respond accordingly.

TABLE VI: Qualitative Comparison of Model Outputs

Generic Prompt	**Competence-Based Instruction**
Recommendations: • Recognition of top-performing regions and employees • Coaching or training of underperforming sales areas • Customer retention strategies targeting high-performing customers • Strategic actions addressing underperforming customers	**Recommendations:** • All measures proposed under the generic prompt, *plus*: • Explicit alignment between sales and marketing on customer needs • Integration of customer data across organisational systems (e.g., CRM, ERP) • Continuous monitoring of industry trends and customer needs • Establishment of formal customer feedback mechanisms

Based on analysed excerpts from model responses `llama3.2:1b,1,1` and `llama3.2:1b,2,3` [38].

The competence-based instruction shifts the model from isolated, performance-driven actions toward coordinated, stakeholder-oriented interventions. Rather than optimising individual sales outcomes, the instructed model proposes measures that integrate multiple organisational functions and explicitly consider long-term customer relationships. This qualitative comparison demonstrates how competence-based prompts operationalise organisational values in concrete decision-making contexts, providing insight beyond numerical evaluation metrics.

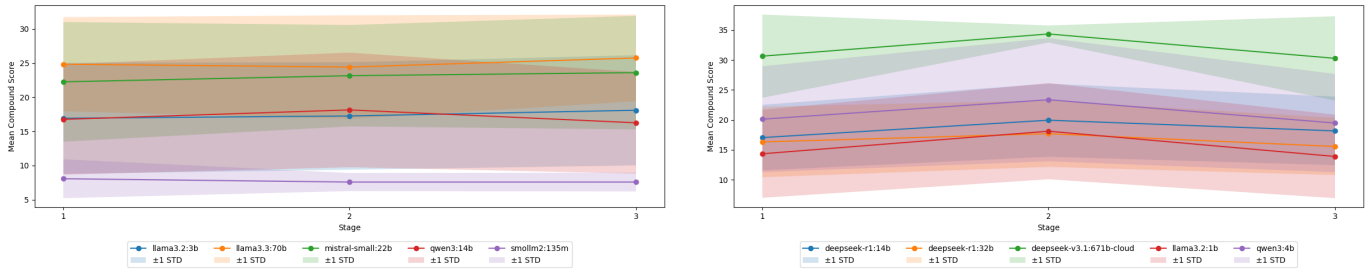

Fig. 2: Mean Compound Score and Standard Deviation Across Stages. Left: Control Group. Right: Treatment Group.

D. Response to Hypotheses

The experimental results provide strong evidence to reject the null hypothesis (H_0) and support the alternative hypothesis (H_1). Competence-based textual instructions significantly improved alignment scores across phases and runs, with Mann–Whitney U tests yielding p-values $< .01$ and rank biserial correlations ranging from .29 to .43, indicating moderate to large effects. The withdrawal stage confirmed reversibility: alignment gains dissipated once instructions were removed, reinforcing causal inference. Textual instructions are the most plausible explanation for the treatment effect because the design controlled for model internals, decoding parameters, and environmental noise, leaving the intervention as the only systematic variable.

V. Discussion

Our study extends prior research on bias mitigation using textual instructions. We contradict prior studies [21] who reported no significant reduction of bias using diversity-promoting instructions, by demonstrating that competence-based prompts achieve measurable improvements. This divergence may stem from the specificity of competence-based instructions, which precisely embed organisational values into textual instructions through behavioural anchors. For SMEs, this distinction is critical: competence-based alignment offers them an effective and scalable method for influencing LLM behaviour.

The experiment highlights evaluator performance as a critical factor in competence-based alignment. M-Prometheus achieved strong agreement with human ratings (Pearson $r = .71$; Cohen's $k = .732$), validating rubric-driven assessment as a reliable proxy for human judgement. Hercule, by contrast, underperformed in English contexts, despite moderate success in German contexts, underscoring the sensitivity of evaluator models to language and rubric translation.

Reference-guided evaluations consistently produced lower alignment scores and reduced inter-rater agreement in the German contextual setting, contradicting prior claims that reference answers enhance evaluation accuracy [26]. This suggests that reference-free approaches may improve adherence to organisational values in non-English contexts, though further research is needed to confirm whether this effect holds true and generalises across domains.

We attribute observed improvements in alignment to competence-based instructions, rather than stochastic variability, supported by a rigorous experimental setup and reversibility in the withdrawal stage. Decoding parameters were deliberately held constant across evaluators to maintain internal validity and isolate the treatment effect [31]. However, this methodological choice introduces a trade-off: uniform configurations may favour certain architectures, while penalising others. M-Prometheus maintained strong correlation with human ratings, whereas Hercule exhibited lower agreement (Pearson $r = .23$ in public phase). Evaluator-specific optimisation may improve fairness and reliability of the results. Prior research confirms that decoding strategies significantly influence output characteristics [4], raising an important question for future work: Should experimental control take precedence over evaluator-specific tuning, or should adaptive strategies be adopted for applied settings?

VI. Limitations and Future Research

While this study demonstrates the potential of competence-based instructions for LLM alignment, several limitations should be considered:

- **Task and domain scope:** The experiment focused on a single sales-analytics task. Generalisation to other tasks, domains, or strategic functions remains to be empirically validated. Future research should *broaden task and domain coverage*, examining areas such as HR, compliance, or customer service.
- **Model sample size and selection:** Only ten LLMs were evaluated. Although the unit of analysis was model output rather than model count, additional architectures may produce different results, particularly for models outside the tested parameter range. We recommend *replicating the experiment using a greater number of models*.
- **Reference answers:** The role of reference answers remains unclear. Reference-guided scoring appears to affect how evaluators interpret rubric criteria, and in our experiment, reference-free scoring resulted in higher agreement. We propose *to investigate the impact of reference answers* in future experiments.
- **Decoding and execution constraints:** Fixed decoding parameters improve internal validity but may disadvantage models with differing optimal configurations, limiting external validity. Subsequent studies should *system-*

atically vary decoding parameters, drawing on research showing that decoding strategies significantly influence outputs [4].

- **SME-specific competence model:** Behavioural anchors were derived from a single organisation, potentially limiting the applicability of the instruction set to other SMEs or cultural contexts. Future research should *test adaptation to other competence models*, taking into account cross-cultural and multilingual settings.
- **Temporal and operational impact:** The study measured alignment as an immediate behavioural outcome. Downstream organisational impact was not assessed. Longitudinal studies should *measure effects on business outcome, decision quality, trust, and strategic cohesion*.

VII. Recommendations for SMEs

We recommend that SMEs take the following measures to align LLMs with OCVG:

- **Develop or adapt a competence model** that reflects your organisation's culture, values and strategic priorities. Express every competence as a behavioural anchor.
- **Design rubrics based on these behavioural anchors** to evaluate LLM outputs, and combine automated scoring with periodic human validation to ensure that automated assessments remain reliable over time.
- **Incorporate the desired competencies into your prompts**, see Chen et al. [20] or our own study for inspiration, and assess their impact through competence-based rubric evaluation [48].

VIII. Conclusion

This study demonstrates that competence-based methods, originally developed for human resource management, can be effectively adapted to align large language models (LLMs) with an organisation's culture, values, and goals (OCVG). Through a randomized controlled trial involving ten open-source LLMs across two datasets and languages, we show that competence-based textual instructions significantly enhance alignment. Treatment models consistently outperformed controls in rubric-driven evaluations, with moderate to large effect sizes.

Our results also reveal that evaluation design and linguistic context matter. Reference-guided assessments produced lower alignment scores and weaker agreement with human ratings in non-English settings, suggesting that reference-free approaches may better preserve normative integrity. While absolute scores were higher using a public English dataset, alignment gains from competence-based instructions were more pronounced when using a SME-specific German dataset. This indicates that linguistic context affects the magnitude of competence-based alignment, and the grades received through competence-based evaluation.

The study contributes a scalable, interpretable strategy for aligning LLMs with organisational values without requiring access to model internals, along with a validated rubric-driven evaluation framework grounded in behavioural anchors. These findings provide empirical evidence that competence models can operationalise organisational values in AI governance, offering SMEs a practical and effective approach to managing the alignment of LLMs with their strategic objectives. Future work should extend these insights across domains, languages, and organisational contexts to further consolidate the link between competence-based alignment and organisational outcomes.

References

[1] McKinsey & Company, "The state of ai in early 2024: Gen ai adoption spikes and starts to generate value," 2024. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

[2] TechRadar/ISACA, "Almost a third of european businesses lack formal ai policies amid generative ai uptake," 2025. [Online]. Available: https://www.techradar.com/pro/security/almost-a-third-of-european-businesses-dont-have-a-formal-comprehensive-ai-policy-in-place-amidst-surging-generative-ai-use-amongst-professionals

[3] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, Sep. 2024, publisher: MIT Press. [Online]. Available: https://doi.org/10.1162/coli_a_00524

[4] A. F. Akyürek, M. Y. Kocyigit, S. Paik, and D. T. Wijaya, "Challenges in Measuring Bias via Open-Ended Language Generation," in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 76–76. [Online]. Available: https://doi.org/10.18653/v1/2022.gebnlp-1.9

[5] R. Chalmers, A. Marras, and G. D. Brannan, "Organizational culture," StatPearls (Internet), Treasure Island (FL): StatPearls Publishing, 2025.

[6] A. T. Bogale, "Organizational culture: A systematic review," *Cogent Social Sciences*, vol. 10, no. 1, p. 2340129, 2024.

[7] A. Vasumathi, A. Vasudevan, A. R. Razak, and S. I. S. Mohammad, "An empirical study on the impact of organizational culture dimensions on employees' performance," *SSAHO*, vol. 10205431, 2025.

[8] N. Gangani, G. N. McLean, and R. A. Braden, "A Competency-Based Human Resource Development Strategy," *Performance Improvement Quarterly*, vol. 19, no. 1, pp. 127–139, Oct. 2008. [Online]. Available: https://doi.org/10.1111/j.1937-8327.2006.tb00361.x

[9] Cameron S. Kim and Quinn E. Robert, *Diagnosing and Changing Organizational Culture : Based on the Competing Values Framework*, third edition ed. San Francisco, CA: Jossey-Bass, Jan. 2011.

[10] R. H. Waterman, T. J. Peters, and J. R. Phillips, "Structure is not organization," *Business Horizons*, vol. 23, no. 3, pp. 14–26, Jun. 1980. [Online]. Available: https://doi.org/10.1016/0007-6813(80)90027-0

[11] R. E. Quinn and J. Rohrbaugh, "A Spatial Model of Effectiveness Criteria: Towards a Competing Values Approach to Organizational Analysis," *Management Science*, vol. 29, no. 3, pp. 363–377, Mar. 1983. [Online]. Available: https://doi.org/10.1287/mnsc.29.3.363

[12] D. R. Denison and A. K. Mishra, "Toward a Theory of Organizational Culture and Effectiveness," *Organization Science*, vol. 6, no. 2, pp. 204–223, Apr. 1995. [Online]. Available: https://doi.org/10.1287/orsc.6.2.204

[13] J. Erpenbeck, L. Rosenstiel, S. Grote, and W. Sauter, *Handbuch Kompetenzmessung : Erkennen, Verstehen und Bewerten Von Kompetenzen in der Betrieblichen, Pädagogischen und Psychologischen Praxis*. Freiburg, GERMANY: Schaffer-Poeschel Verlag fur Wirtschaft Steuern Recht GmbH, 2017.

[14] R. Sanchez, "Understanding competence-based management: Identifying and managing five modes of competence," *Journal of Business Research*, vol. 57, no. 5, pp. 518–532, 2004.

[15] I. Klepić, "The Influence of Human Resources Competency Management on the Business Success of Small and Medium Enterprises." *Our Economy / Nase Gospodarstvo*, vol. 68, no. 4, pp. 12–27, 2022. [Online]. Available: https://doi.org/10.2478/ngoe-2022-0020

[16] P. Cardona and C. Rey, *Management by Missions: Connecting People to Strategy through Purpose*. Cham: Springer International Publishing, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-83780-8

[17] A. Moreira, M. J. Sousa, and F. Cesário, "Competencies Development: The Role of Organizational Commitment and the Perception of Employability." *Social Sciences (2076-0760)*, vol. 11, no. 3, p. 125, Mar. 2022, publisher: MDPI. [Online]. Available: https://doi.org/10.3390/socsci11030125

[18] D. Horvat, A. Jäger, and C. M. Lerch, "Fostering innovation by complementing human competences and emerging technologies: an industry 5.0 perspective." *International Journal of Production Research*, vol. 63, no. 3, pp. 1126–1149, 2025. [Online]. Available: https://doi.org/10.1080/00207543.2024.2372009

[19] S. Vilches, S. Fenwick, B. Harris, B. Lammi, and R. Racette, "Changing health organizations with the LEADS leadership framework: Report of the 2014-2016 LEADS impact study." Fenwick Leadership Explorations, the Canadian College of Health Leaders, & the Centre for Health Leadership and Research, Royal Roads University, Tech. Rep., 2016.

[20] Y. Chen, V. C. Raghuram, J. Mattern, R. Mihalcea, and Z. Jin, "Causally Testing Gender Bias in LLMs: A Case Study on Occupational Bias," in *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, 2025, pp. 4984–5004. [Online]. Available: https://doi.org/10.18653/v1/2025.findings-naacl.281

[21] C. Borchers, D. Gala, B. Gilburt, E. Oravkin, W. Bounsi, Y. M. Asano, and H. Kirk, "Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements," in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 212–224. [Online]. Available: https://doi.org/10.18653/v1/2022.gebnlp-1.22

[22] Anthropic, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2023. [Online]. Available: https://arxiv.org/abs/2212.08073

[23] NVIDIA Corporation, "Nemo guardrails: Toolkit for controlling and safeguarding llm applications," https://github.com/NVIDIA/NeMo-Guardrails, 2023, accessed: 2026-04-04.

[24] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 46 595–46 623.

[25] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 150–158. [Online]. Available: https://doi.org/10.18653/v1/2024.eacl-demo.16

[26] S. Kim, J. Shin, y. cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, Ryan, S. Kim, J. Thorne, and M. Seo, "Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models," in *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 29 927–29 962.

[27] S. Doddapaneni, M. S. U. R. Khan, D. Venkatesh, R. Dabre, A. Kunchukuttan, and M. M. Khapra, "Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 29 297–29 329. [Online]. Available: https://doi.org/10.18653/v1/2025.acl-long.1419

[28] J. Pombal, D. Yoon, P. Fernandes, I. Wu, S. Kim, R. Rei, G. Neubig, and A. F. T. Martins, "M-Prometheus: A Suite of Open Multilingual LLM Judges," 2025, version Number: 1. [Online]. Available: https://doi.org/10.48550/ARXIV.2504.04953

[29] J. Park, S. Jwa, R. Meiying, D. Kim, and S. Choi, "OffsetBias: Leveraging Debiased Data for Tuning Evaluators," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 1043–1067. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-emnlp.57

[30] K. Wataoka, T. Takahashi, and R. Ri, "Self-Preference Bias in LLM-as-a-Judge," 2024, version Number: 2. [Online]. Available: https://doi.org/10.48550/ARXIV.2410.21819

[31] A. E. Kazdin, *Research design in clinical psychology*, fifth edition ed. Boston Munich: Pearson, 2017.

[32] Pandas, "pandas.DataFrame.sample — pandas 3.0.0 documentation." [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sample.html

[33] W. G. Cochran, *Sampling techniques*, 3rd ed., ser. Wiley series in probability and mathematical statistics. New York, NY: Wiley, 1977.

[34] Daqing Chen, "Online Retail II," 2012. [Online]. Available: https://archive.ics.uci.edu/dataset/502

[35] Ollama, "Ollama," 2025. [Online]. Available: https://ollama.com/

[36] ——, "Cloud · Ollama," 2025. [Online]. Available: https://ollama.com/cloud

[37] vLLM, "Using Docker - vLLM," Jul. 2025. [Online]. Available: https://docs.vllm.ai/en/stable/deployment/docker.html

[38] "Prof-it/ai-alignment: Thie repo keeps the code experimenting on the AI-alignment to the company value, based on a thesis." [Online]. Available: https://github.com/Prof-it/ai-alignment/blob/main/experiment/score_rubrics_en.json

[39] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951. [Online]. Available: https://doi.org/10.1007/BF02310555

[40] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, Dec. 1965. [Online]. Available: https://doi.org/10.1093/biomet/52.3-4.591

[41] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, Mar. 1947. [Online]. Available: http://doi.org/10.1214/aoms/1177730491

[42] G. V. Glass, "Note on Rank Biserial Correlation," *Educational and Psychological Measurement*, vol. 26, no. 3, pp. 623–631, Oct. 1966. [Online]. Available: https://doi.org/10.1177/001316446602600307

[43] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. New York, NY: Psychology Press, 2009.

[44] K. Pearson, "VII. Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240–242, Dec. 1895. [Online]. Available: https://doi.org/10.1098/rspl.1895.0041

[45] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, no. 1, p. 72, Jan. 1904. [Online]. Available: https://doi.org/10.2307/1412159

[46] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[47] G. E. Gignac and E. T. Szodorai, "Effect size guidelines for individual differences researchers," *Personality and Individual Differences*, vol. 102, pp. 74–78, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.paid.2016.06.069

[48] W. Sauter and F.-P. Staudt, "Kompetenzmodelle," in *Strategisches Kompetenzmanagement 2.0*. Wiesbaden: Springer Fachmedien Wiesbaden, 2016, pp. 9–19, series Title: essentials. [Online]. Available: http://doi.org/10.1007/978-3-658-11294-3_2

IX. Appendix

A. Use of artificial intelligence

Artificial intelligence (Microsoft Copilot) tools were used to support the writing process; however, all referenced sources were collected and thoroughly checked by the authors, and the experiment—including data preparation, model execution, and evaluation—was fully conducted by the authors.

B. LLM-as-a-Judge Deployment

LLM-as-a-Judge systems were central to our evaluation methodology. Open-source evaluators M-Prometheus and Hercule scored outputs against competence-based rubrics, with 8.73% manually checked to benchmark human agreement. The fully automated, containerised setup ensured reproducibility, scalability, and transparency; detailed configurations are available in the repository.