International Journal of Advance Research in Science, Management, Engineering and Technology
Volume 1, Issue 5 | November - December 2026 | www.ijarsmet.com
ISSN: **3139-1141**

A Study of the Effects of Language on Large Language Model Responses and Human Perception

Shubham Gunjal[#1], Siddhesh Ghodekar[#2], Dr. Rupali Pawar[*3]

[#]*Department of Master of Computer Applications (Semester III), Zeal Institute of Business Administration, Computer Application and Research (ZIBACAR),*

Savitribai Phule Pune University, Pune, India

[1]gunjalshubham89@gmail.com , [2]ghodekarsiddhesh19@gmail.com , [3]Rupali.pawar@zealeducation.com

Abstract—

People often feel an idea more strongly in their mother tongue, and they soften a hard message so that the listener is not hurt. This study asks whether chat-based large language models (LLMs) handle hard messages with the same care in English, Hindi and Marathi. Ten industry situations, such as a rejected build, a loan denial and the removal of an engineer from duty, were given to ChatGPT, Claude and Gemini in the three languages. The 90 collected messages were scored with a concept-matched word list that has five harshness categories and four softening categories. The scores were Harshness H, Softening S and Net tone N = H − S, with content coverage and length as extra checks. Differences between models were large (Kendall's W of 0.30 for N and 0.56 for S). Differences between languages were small and not significant after correction. Gemini wrote short messages with almost no softening (S = 0.06), whereas ChatGPT and Claude used courtesy, regret and offers of help. Gemini also left out more key facts in Marathi (82% coverage) than in
English (94%). Each cell holds one response, the word list was built after reading the data, and no human rater was used, so the results are exploratory.

Index Terms—Large language models, multilingual NLP, politeness, softening, Hindi, Marathi, tone measurement, human perception.

I. Introduction

Many people now use chat assistants to write difficult messages. A manager may need to reject a build. A hotel may have to refuse a refund. A university may have to cancel an exam paper. In each case the writer must say something unpleasant without breaking the relationship with the reader.

Humans handle this problem all the time. We choose our words, our tone and sometimes even our language. Most of us talk to family in our mother tongue. In that language an idea, a joke or a rebuke often feels heavier than the same idea in a language learned at school. Because of this, speakers sometimes move to another language, or to a gentler version of the same sentence, when they want to be clear but not cruel.

This raises two questions. The first is whether such softening works in daily life, that is, whether it makes talking more effective. The second is whether today's AI models use the same habit. A model that writes a Hindi rejection may be as careful as one that writes it in English, or it may be harsher, or polite only on the surface. If the second case is true, users in different languages would receive different levels of care from the same system.

This paper takes a first step towards both questions with a small but real data set. The author asked three chat models to write the same ten difficult workplace messages in English, Hindi and Marathi and collected the 90 responses. The paper then measures how harsh and how soft each message is, how much of the key content it keeps, and how long it is. All calculations are described in Section VI, so that anyone can repeat them. The human side of the question, that is how real readers feel these messages, is not tested here and is left as the next step.

II. Background and Motivation

Language is tied to experience. A first language is usually learned at home, together with family habits, jokes and quarrels. A later language is often learned in a classroom and used for work. Studies of bilingual people report that the first language tends to carry more emotional weight, while a later language can feel more distant [3, 4]. This distance has measurable effects. People can make more even-handed decisions in a foreign language [1] and can judge a moral dilemma differently [2].

Everyday speech uses a second tool as well, namely softening. A direct sentence such as "Your work is poor" can be changed to "Some parts of your work can be improved." The facts are close, but the feeling is not. Politeness theory describes such changes as ways of protecting the listener's face when a message threatens it [5].

Chat models are trained on text from many languages, but the amount of text differs from one language to another. English is far larger than Hindi, and Marathi is smaller still. It is therefore reasonable to ask whether care in wording is equally good in all three. The question matters in India, where one organisation may send the same notice in several languages on the same day.

III. Related Work

A. Language and Emotion in People

Keysar et al. found that decisions made in a foreign language were less affected by framing and loss feelings than decisions

ISSN: **3139-1141**

made in the native language [1]. Costa et al. reported that people gave different answers to a moral dilemma when it was read in a foreign language [2]. Caldwell-Harris reviewed the evidence and described the native language as the one in which emotional words, reprimands and childhood memories usually have the strongest effect [3]. Pavlenko reached a similar view from the study of bilingual speakers [4]. These studies are about human speakers. They give a reason to expect that wording and language together change how a hard message is felt, but they do not test AI systems.

B. Politeness and Softening

Brown and Levinson described how speakers use indirect forms, hedges and respect words to reduce the threat of an unpleasant act [5]. In language technology, Madaan et al. created the task of politeness transfer, in which a rude sentence is rewritten as a polite one while its content is kept [6]. Their work shows that softening can be measured with two goals at the same time: the style should change and the meaning should stay. This paper follows the same two-part idea by measuring tone and content coverage separately.

C. Language Models, Culture and Politeness

Tao et al. compared model answers with survey data from many countries and found that the answers were closer to the values of English-speaking countries [9]. Hu et al. compared people and models on seven pragmatic tasks and found that models were weaker where social expectations were broken [10]. Closer to this paper, Yin et al. tested how the politeness of a *prompt* affects model performance in English, Chinese and Japanese, and found that the best level changed with the language [7]. Mehta et al. extended this line to English, Hindi and Spanish with five models and reported that the best tone depended on the language and the model [8]. Both studies look at the politeness of the input and its effect on answer quality. This paper looks at the tone of the *output* when the task itself is to deliver bad news.

D. Tools for Measuring Meaning and Tone

To compare the meaning of texts in different languages, a multilingual sentence embedding is a common choice [11, 12, 13]. For Marathi, labelled sentiment data exist [14], but tone and politeness tools for Hindi and Marathi are far fewer than for English. For this reasion the present study uses a transparent word-list method, which can be checked by hand, and states its limits in Section X.

IV. Research Gap and Objectives

The literature shows that language affects emotion in people, that politeness can be studied by computer, and that the politeness of a prompt can change the quality of an answer. In the searches made for this study, no work was found that compares how ChatGPT, Claude and Gemini write the *same hard message* in English, Hindi and Marathi. The objectives are:

1. To collect responses of three chat models to ten sensitive scenarios in three languages.
2. To define simple, repeatable measures of harshness, softening, content coverage and length.
3. To test whether the measures differ between languages and between models.

TABLE I
The ten scenarios

ID	Industry	Hard message
S1	IT and software QA	Build rejected after failed audits
S2	Healthcare administration	Doctor two hours late, patients at risk
S3	Banking and mortgage	Loan declined, documents do not match
S4	Automotive assembly	Line stopped, brake failure found
S5	E-commerce logistics	Vendor contract terminated
S6	Corporate legal and IP	Campaign withdrawn, trademark conflict
S7	Luxury hotel management	Refund and upgrade refused, damage to property
S8	University examination	Exam cancelled, unauthorised notes
S9	Civil construction	Work rejected, cement grade and cracks
S10	Airline operations	Engineer removed from duty warning ignored

,

4. To check how much the findings depend on the choices made in the word list.
5. To prepare the human rating study that the original synopsis proposes.

V. Data and Scenario Design

A. Scenarios

Ten scenarios were prepared, one for each industry (Table I). Each scenario is a situation in which the writer has to deliver a hard decision, for example a failed audit, a patient put at risk, a refused loan or a cancelled exam.

B. Collection of Responses

The author gave each scenario to ChatGPT, Claude and Gemini by hand through their chat interfaces, once in English, once in Hindi and once in Marathi. This gives $10 \times 3 \times 3 = 90$ responses, with one response in each cell. The responses were pasted into a text file without editing. Claude's responses are written as a full e-mail (subject line, greeting, closing and placeholders such as the name of the sender), while ChatGPT and Gemini wrote only the message body. The model versions, the collection dates and the exact prompt wording were not stored in the file and are therefore not reported here.

VI. Methodology

Fig. 1 shows the workflow. The ten scenarios become requests in three languages, the requests are given to three models, and the 90 responses are scored. Statistical tests and sensitivity checks follow. The last box, a rating study with human readers, is planned and was not carried out.

A. Preparing the Text

Each response is split into a *body* and a *frame*. The frame holds the subject line, the greeting, the closing and the signature. Only the body is scored, so that Claude's e-mail format does not change its tone scores. The frame is reported separately. Hindi and Marathi text is normalised (Unicode NFC, dots under letters removed, Devanagari digits changed to ordinary digits, chandrabindu changed to anusvara) and English text is lower-cased. Sentences are split at full stops, question marks, exclamation marks and the Devanagari danda, with common abbreviations protected. A word is any space-separated piece of text that contains a letter or digit.

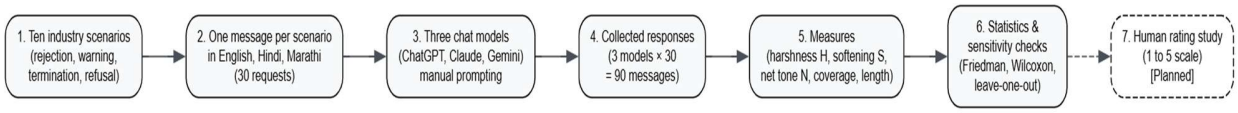

Fig. 1. Workflow of the study. Solid boxes were completed. The dashed box is the planned human rating study.

TABLE II TONE CATEGORIES (ENGLISH EXAMPLES)

Code	Idea	Examples
URG	Urgency, deadline	immediately, within 48 hours
ABS	Absolute or final wording	final, non-negotiable, strictly
CMD	Order, obligation	must, required, do not
THR	Penalty, legal or disciplinary step	legal action, disciplinary, charges
CERT	Accusation stated as fact	fraud, harassment, confirmed
COURT	Courtesy	please, thank you, appreciate
REGRET	Regret, empathy	regret, we understand, unfortunately
HEDGE	Cautious or indirect wording	may, could, at this time, appears
OFFER	Offer of help or discussion	available, discuss, opportunity

B. Word List for Tone

A tone word list was prepared with the same concepts in all three languages, so that each language is checked for the same idea and not for the same English word. Table II describes the nine categories. Hindi and Marathi lists were written in Devanagari script (for example *kripya*, *dhanyavad*, *sakte hain* and *shakto* in Roman script). A category is *present* in a message if at least one of its patterns appears in the body.

CERT is counted per sentence: a sentence counts when it contains an accusation word and no hedge word, because "the documents appear to be inconsistent" is gentler than "the documents are fraudulent."

C. How Each Measure Is Calculated

Let $c_k = 1$ if category k is present in a message and 0 otherwise. For one message:

$$H = \tfrac{1}{5}\,(c_{URG} + c_{ABS} + c_{CMD} + c_{THR} + c_{CERT}), \quad (1)$$

$$S = \tfrac{1}{4}\,(c_{COURT} + c_{REGRET} + c_{HEDGE} + c_{OFFER}), \quad (2)$$

$$N = H - S. \quad (3)$$

H and S lie between 0 and 1, and N between −1 and +1. A positive N means that the message is blunter than it is soft. Because a category counts once however often it appears, H and S measure the *variety* of tone devices. To capture how often they are used, the marker counts per 100 words are also given:

$$M_H = 100 \ \frac{\text{harsh hits}}{\text{words}}, \qquad M_S = 100 \ \frac{\text{soft hits}}{\text{words}}. \quad (4)$$

Harshness here is a property of the wording and not a judgement of the decision. A loan refusal must contain a hard fact, so some harshness is unavoidable.

Content coverage (COV) checks that a message still says what the scenario needs. For every scenario five key elements were fixed (for example, for S4: stop the line, brake failure, isolate the units, find the root cause, no dispatch). Each element has one pattern in each language. COV is the share of the five elements found in the message, so it lies between 0 and 1. There are 147 elements in all, which gives 441 checks over the 90 messages. Lastly, *length* is measured in words and sentences of the body.

D. Statistics

Each scenario was answered by every model in every language, so the data are paired. For the language effect the 30 blocks are the combinations of model and scenario. For the model effect the 30 blocks are the combinations of language and scenario. A Friedman test [15] checks the difference across the three levels, and Kendall's $W = \chi^2/(n(k-1))$ gives its size. Pairwise Wilcoxon signed-rank tests [16] follow, with Holm correction [17]. The blocks are not fully independent, because the same scenario appears in several blocks, so p values are a guide and not proof. To check the word list, each of the nine categories was left out in turn and N was recalculated.

VII. EXPERIMENTAL SETUP

The analysis used Python with Pandas, NumPy, SciPy and Matplotlib, and the workflow figure was drawn with Graphviz. No random numbers were used, so there is no seed to report.

Every number in this paper can be recalculated from the 90 stored responses and the word list. The word list was written after reading the responses. It was then refined in two rounds to remove false matches, such as the word "required" in "as required" and nouns that look like verbs in Hindi. This makes the method partly tuned to this data, which is a limitation (Section X).

VIII. Results

A. Overall Scores

Table III gives the means for each model and language. Across all 90 messages the mean harshness is $H = 0.46$, the mean softening is $S = 0.28$ and the mean net tone is $N = 0.18$. Claude has the highest harshness (0.55) and Gemini has by far the lowest softening (0.06). ChatGPT is the most balanced

TABLE III
MEAN SCORES BY MODEL AND LANGUAGE

Model	Lang.	H	S	N	M_H	M_S	Words	COV
ChatGPT	EN	0.38	0.45	−0.07	2.89	2.59	97.9	1.00
	HI	0.38	0.32	0.06	2.55	2.01	114.9	0.98
	MR	0.36	0.28	0.08	3.23	2.37	87.5	0.96
Claude	EN	0.54	0.42	0.12	4.40	2.86	84.7	1.00
	HI	0.54	0.42	0.12	4.71	2.58	93.2	1.00
	MR	0.58	0.40	0.18	6.00	2.90	72.8	1.00
Gemini	EN	0.48	0.05	0.43	6.57	0.38	49.7	0.94
	HI	0.40	0.05	0.35	5.67	0.33	54.9	0.92
	MR	0.44	0.08	0.37	6.88	0.68	43.9	0.82
All		0.46	0.28	0.18	4.77	1.85	77.7	0.96

M_H, M_S: harsh and soft hits per 100 words. COV: content coverage. Each row has 10 messages.

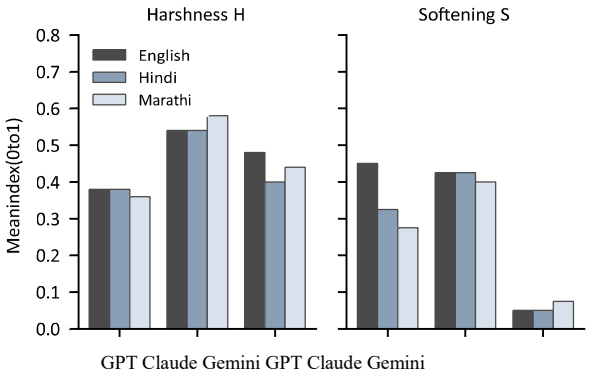

Fig. 2. Mean harshness H and softening S by model and language.

($N = 0.02$). Fig. 2 shows H and S side by side.

B. Differences Between Languages

For harshness, the language effect is very small (Friedman $\chi^2 = 1.08$, $p = 0.58$, $W = 0.02$). The mean H is 0.47 in English, 0.44 in Hindi and 0.46 in Marathi. Softening drops from 0.31 in English to 0.27 in Hindi and 0.25 in Marathi, and the overall test is significant ($\chi^2 = 7.23$, $p = 0.027$, $W = 0.12$), but no single pair survives the Holm correction (English against Marathi: corrected $p = 0.059$). The net tone rises slightly from English (0.16) to Hindi (0.17) and Marathi (0.21), which is not significant ($p = 0.29$). The Marathi value is the highest in all ten leave-one-category-out versions of N (Section G), so the direction is stable, even though the size is small.

The drop in softening is clearest in courtesy words. Courtesy was present in 57% of English messages, but in 40% of Hindi and 37% of Marathi messages. For ChatGPT, the courtesy rate was 60%, 20% and 10% in English, Hindi and Marathi. This may be a real habit of the model, or it may be that Hindi and Marathi messages express courtesy through verb forms that a word list cannot see. Human readers are needed to decide.

The word count differs strongly between languages ($W = 0.92$). Hindi messages are about 10 words longer than English ones and Marathi messages about 9 words shorter. This is mainly a feature of the languages, because Hindi writes small function

TABLE IV
STATISTICAL TESTS (FRIEDMAN, n=30 BLOCKS)

Factor	Measure	χ^2	p	W
Language	H			
	S			
	N			
	Words	55.27		
Model	H	1.08	0.584	0.02
	S	7.23	0.027	0.12
	N	2.49	0.288	0.04
	Words		<0.001	0.92
		19.08	<0.001	0.32
		33.74	<0.001	0.56
		17.96	<0.001	0.30
			<0.001	0.91
		54.60		

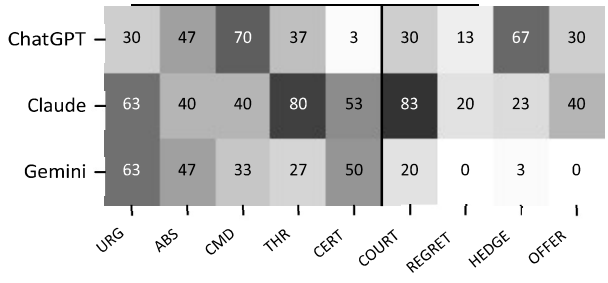

Fig. 3. Share of messages (%) in which each category is present. The first five columns are harshness categories and the last four are softening categories.

words such as *ke* and *hai* separately, whereas Marathi attaches them to the word. Word counts, and therefore the per-100-word

rates, should not be compared across languages as signs of effort or tone.

C. Differences Between Models

The model effect is large for all three tone scores (Table IV). Pairwise tests with Holm correction give the following. Claude is harsher than ChatGPT (mean difference 0.18, corrected $p < 0.001$) and Gemini (0.11, $p = 0.012$). Gemini is much less soft than both ChatGPT (0.29) and Claude (0.36), with $p < 0.001$ in both cases. Net tone orders the models clearly: Gemini > Claude > ChatGPT, with corrected p values of 0.002 (Gemini and Claude), 0.049 (Claude and ChatGPT) and <0.001 (Gemini and ChatGPT).

Fig. 3 shows which categories drive this. ChatGPT uses obligation words (70%) and cautious words (67%) but rarely states accusations as fact (3%). Claude combines courtesy (83%) with consequence words (80%) and urgency (63%). Gemini writes short, firm messages: 63% urgency, 50% accusations stated as fact, and almost no courtesy (20%), regret (0%), cautious wording (3%) or offer of help (0%). The length data agree: Gemini used 2.8 sentences and 49.5 words on average, against 5.6 and 100.1 for ChatGPT and 5.8 and 83.6 for Claude.

Harshness and softening are not related to each other across messages (Spearman $\rho = -0.01$, $p = 0.91$). A message can therefore be both firm and courteous, as Claude's often are. Softening is related to length ($\rho = 0.53$, $p < 0.001$): longer messages have more room for softening, whereas harshness is not related to length ($\rho = -0.11$, $p = 0.31$).

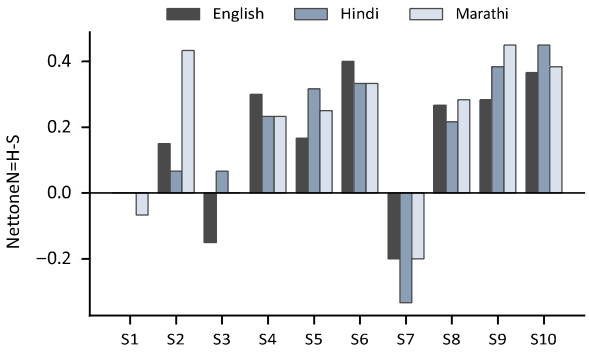

Fig. 4. Net tone N by scenario and language (mean of three models).

D. Consistency Across Languages

For each model and scenario, the spread of N across the three languages was calculated (largest value minus smallest). The mean spread was 0.09 for Claude, 0.23 for Gemini and 0.28 for ChatGPT. Claude's three versions of a message are therefore nearly the same in tone, which is what a faithful translation would give. They also have the same number of sentences (5.8) in each language. ChatGPT changes its tone more from one language to another, mainly through the courtesy and offer categories. The mean absolute difference of N between two languages was 0.13 for all three pairs (English–Hindi 0.128, English–Marathi 0.135, Hindi–Marathi 0.133).

E. Scenario Effects

The scenario matters as much as the language (Fig. 4). Harshness was highest for the legal case S6 (0.69) and lowest for the build rejection S1 (0.31). Net tone was negative in the hotel refusal S7 (mean −0.24), where the models offered much help and courtesy (mean $S = 0.67$). It was close to zero for the build rejection S1 and the loan denial S3, and highest for the airline case S10 (0.40) and construction case S9 (0.37), where safety is involved. The models thus change their tone with the situation, in the same direction in all three languages for most scenarios.

F. Content Coverage

Of the 441 element checks, 95.7% were found (English 98.0%, Hindi 96.6%, Marathi 92.5%). Claude kept all elements in all languages. ChatGPT scored 100%, 98% and 96% and Gemini 94%, 92% and 82%. Of the 147 elements, 10 (6.8%) were present in one language and missing in another, and in all 10 cases the missing one was in Hindi or Marathi, mostly Marathi (Table V). An example is Gemini's Marathi message for S1, which does not say that the audit failed. Because coverage uses patterns too, some of these gaps may be differences in wording and not missing facts, so each should be checked by a bilingual reader.

G. Sensitivity of the Findings

Leaving out each category in turn gave ten versions of N (including the full version). Gemini had the highest net tone in all ten. Claude was higher than ChatGPT in nine of the ten, and the exception was the version without the hedge category (0.08 against 0.13). Marathi had the highest net tone among the languages in all ten, and English the lowest in nine, the exception being the version without courtesy, where Hindi was lowest (0.22 against 0.24). The model ranking is therefore stable, and the small language ranking is also stable but its size is small and not significant.

TABLE V
ELEMENTS PRESENT IN ENGLISH BUT MISSING IN ANOTHER LANGUAGE

Model	Scenario: element	Missing in
ChatGPT	S4: no dispatch	MR
ChatGPT	S6: withdraw campaign	HI
ChatGPT	S10: removed from duty	MR
Gemini	S1: audit failed; rewrite	MR
Gemini	S4: isolate units; root cause	MR
Gemini	S5: handover	MR
Gemini	S8: chance to reply	HI
Gemini	S10: removed from duty	MR

IX. DISCUSSION

A. Answers to the Research Questions

The first question was whether language changes how AI writes a hard message. In this data the language has only a small effect on tone: harshness did not differ, and softening was a little lower in Marathi and Hindi, mostly because fewer courtesy words

ISSN: **3139-1141**

were found. The choice of *model* had a much larger effect than the choice of language, and the scenario also mattered a lot. The second question was whether the models soften their messages in the way people do. ChatGPT and Claude did, in all three languages, by using courtesy, regret and offers to talk. Gemini mostly did not. Its messages were short and firm in all three languages, so for Gemini the missing softening is a model habit and not a language problem. The one place where Gemini was clearly worse in an Indian language was content: in Marathi it left out more of the key facts.

The original idea of the study, that a mother tongue carries stronger feeling and that softened wording can protect the reader, concerns how *people* receive messages. This paper cannot say whether softened messages work better in real life. That needs readers.

B. What This Means in Practice

A user who asks for a hard message in Hindi or Marathi can expect about the same level of harshness as in English. They should not expect the same level of courtesy from every model, and they should read Marathi output from Gemini with care, because some facts may be missing.

C. Plan for the Next Step

Bilingual volunteers should read a sample of the 90 messages in their own language and rate harshness, politeness, discomfort, clarity, trust and willingness to accept on a 1 to 5 scale. Their ratings can be compared with *H*, *S* and *N* to see whether the word list agrees with people. A multilingual sentence embedding [11, 13] can then replace the content-coverage patterns. Each prompt should be repeated several times, because a model can give different answers to the same request.

X. Limitations and Ethical Considerations *A.*

Limitations

First, each cell holds one response, so random variation inside a model is not measured. Second, the word list was built after reading these responses. Its results are therefore partly tuned to the data, and the leave-one-out check reduces but does not remove this problem. Third, a word list sees words and not meaning. Hindi and Marathi express courtesy and doubt through verb endings, which are hard to match, so *S* may be too low in those languages. Fourth, *H* and *S* show whether a category is present and not how strong it is. Fifth, the prompts, model versions and dates were not stored. Sixth, only ten scenarios and three models were used. Seventh, no human rater took part, so the scores show differences in wording and not how readers feel. Finally, the Friedman tests treat blocks as independent although scenarios repeat.

B. Ethical Considerations

The planned human study will show workplace messages to volunteers. Participation should be voluntary, with informed consent and no personal data in the messages. Softening has a second side as well: it can hide a serious problem, so a softened message should still state the real issue. Finally, tone tools that work well in English may be unfair to speakers of other languages, so any score used in decisions about people must be tested in each language first.

XI. Conclusion

This study collected 90 real responses of ChatGPT, Claude and Gemini to ten hard workplace scenarios in English, Hindi and Marathi and measured their harshness, softening, net tone, content coverage and length with a transparent word list. The model made a large difference: Claude was the harshest, Gemini was the least soft and the shortest, and ChatGPT was the most balanced. The language made only a small difference in tone, with slightly less softening in Marathi and Hindi, which was not significant after correction. The clearest sign of a language gap was that Gemini left out more key facts in Marathi. The findings are exploratory because of single responses, a word list built after reading the data, and the lack of human raters. The next step is a human rating study, to show whether people feel the same differences that the word list finds.

References

[1] B. Keysar, S. L. Hayakawa, and S. G. An, “The foreignlanguage effect: Thinking in a foreign tongue reduces decision biases,” *Psychological Science*, vol. 23, no. 6, pp. 661–668, 2012.

[2] A. Costa, A. Foucart, S. Hayakawa, M. Aparici, J. Apesteguia, J. Heafner, and B. Keysar, “Your morals depend on language,” *PLoS ONE*, vol. 9, no. 4, p. e94842, 2014.

[3] C. L. Caldwell-Harris, “Emotionality differences between a native and foreign language: Theoretical implications,” *Frontiers in Psychology*, vol. 5, art. 1055, 2014.

[4] A. Pavlenko, “Affective processing in bilingual speakers: Disembodied cognition”, *International Journal of Psychology*, vol. 47, no. 6, pp. 405–428, 2012.

[5] P. Brown and S. C. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[6] A. Madaan *et al.*, “Politeness transfer: A tag and generate approach,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020.

[7] Z. Yin, H. Wang, K. Horio, D. Kawahara, and S. Sekine, “Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance,” arXiv preprint arXiv:2402.14531, 2024.

[8] H. Mehta, A. Saxena, G. Chhikara, and R. Kumar, “No universal courtesy: A cross-linguistic, multi-model study of politeness effects on LLMs using the PLUM corpus,” arXiv preprint arXiv:2604.16275, 2026.

ISSN: **3139-1141**

[9] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, no. 9, p. pgae346, 2024.

[10] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson, "A fine-grained comparison of pragmatic language understanding in humans and language models," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 4194–4213.

[11] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2020, pp. 4512–4525.

[12] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020.

[13] S. Khanuja *et al.*, "MuRIL: Multilingual representations for Indian languages," arXiv preprint arXiv:2103.10730, 2021.

[14] A. Kulkarni *et al.*, "L3CubeMahaSent: A Marathi tweetbased sentiment analysis dataset," in *Proc. 11th Workshop Computational Approaches Subjectivity, Sentiment Social media Anal. (WASSA)*, 2021.

[15] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.

[16] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[17] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.