Investigating LLM's Knowledge about English G2P Rules and Pronunciation with Pseudo-words

by

Sheng Yao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2026

© Sheng Yao 2026

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

For large language models (LLM), the real world is mapped onto a world made of text strings, and one current research direction in NLP is to examine how much knowledge LLMs learn inside their text world. Studies have shown that they have knowledge of English grapheme-to-phoneme conversion (G2P) and pronunciation, but only to a moderate degree. They mainly use tasks such as rhyme detection, syllable counting, and G2P involving words inside the vocabulary of a language - in most cases English. While the results are convincing, we still believe that the acid test of such knowledge should involve pseudo-words - made-up orthographic words. For example, state-of-the-art LLMs such as GPT5 and Gemini3-Pro have no problem providing the pronunciation of an in-vocabulary English word as they are clever enough to fetch the fact from their training data. When given a pseudo-word, however, their predictions can sound unnatural from a human perspective. On the other hand, if human participants all agree on a certain pronunciation for a given pseudo-word, it means they have used some common (implicit) knowledge about G2P and pronunciation when making their prediction. Therefore, we aim to examine the degree of similarity between human participants and LLMs when they are predicting the sound of pseudo-words as an indicator of whether LLMs have learned about G2P and pronunciation in their text world.

It turns out that LLMs' knowledge does have a remarkable degree of human-likeness, not only because 80% of LLMs' predictions are the same as humans' when there is zero inter-human divergence, but also because LLMs' bewilderment (measured by how LLMs' predictions vary across runs) correlates with humans'. That is, models and humans are dealing with the G2P task in more or less the same way.

However, we also see substantial numbers of cases where LLMs' predictions are far from humans' predictions. When we took a closer look at such cases, we found a couple of tendencies and further validated the findings using real English words. More than half of these tendencies suggest that LLMs actually oversimplify the matter of G2P, sticking to the most common mappings in the English vocabulary. These tendencies can be useful when we try to improve the performance of LLMs and even text-to-speech models on pronunciation tasks.

In fact, we also included the text-to-speech component of SpeechT5, in order to compare the performance of text-only models and bi-modal ones. We find that, while there seems to be a bottleneck for LLMs in the sense that the most powerful model, GPT5.4, does not significantly outperform a much weaker Llama3 on the G2P task, SpeechT5 is easily more human-like than all LLMs on several metrics. It seems that bi-model learning does give text-to-speech models such as SpeechT5 an advantage on a sound-related task, despite the fact that SpeechT5 is much smaller than LLMs.

Acknowledgements

I would like to thank my supervisor Professor Freda Shi for supporting me over the two years. Her guidance on research always proves the most helpful when I feel the most helpless.

Also great thanks to Professor Randy Harris and Professor Robin Cohen for being on the thesis committee and providing me with valuable feedback.

Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures vii

List of Tables viii

1 Introduction 1
- 1.1 Background and Our Contribution . . . 1
- 1.2 Related Works . . . 3
  - 1.2.1 Linguistic Knowledge of LLMs . . . 3
  - 1.2.2 Humans on Pseudo-word Tasks . . . 4

2 Experiment Settings 5
- 2.1 Generating Pseudo-Words . . . 5
- 2.2 Human Participants . . . 8
- 2.3 Models . . . 8
- 2.4 Model Inference . . . 10
- 2.5 Transcription . . . 11
- 2.6 Computing the Editing Distance . . . 14

3 Analysing the Performance 16
- 3.1 Metrics and Methods . . . 17
- 3.2 Human-likeness in Terms of Performance . . . 20
- 3.3 Human-likeness in Terms of Closeness . . . 22
  - 3.3.1 The Model-Human Distance . . . 22
  - 3.3.2 The Closest Model-Human Distance . . . 24
  - 3.3.3 The Earth Mover's Distance (EMD) . . . 26
- 3.4 Human-likeness in Terms of Feeling the Difficulty . . . 36
- 3.5 Discussion . . . 41
  - 3.5.1 Human-likeness . . . 41
  - 3.5.2 Comparing Models . . . 41

4 Analysing Tendencies 43
- 4.1 Tendencies with Consonants . . . 44
  - 4.1.1 Rigid Mappings to Fricatives . . . 44
  - 4.1.2 Failure to Handle Silent Consonants . . . 46
  - 4.1.3 Extra [j] Mapped from 'u' . . . 47

4.2 Tendencies with Vowels . . . 48
4.2.1 Vertical Shifting of Monophthongs . . . 52
4.2.2 Shortened Glide of Diphthongs . . . 56
4.2.3 Influence of String Length . . . 57
4.2.4 Failure to Handle the Silent 'e' . . . 58
4.2.5 Extra [ə] in Rounded Vowels . . . 59
4.2.6 Failure to Centralise Unstressed Vowels . . . 60
4.3 Discussion . . . 60

5 Conclusion and Discussion **62**
5.1 Summary of Contributions . . . 62
5.2 Limitations . . . 62
5.3 Future Directions . . . 63
5.3.1 Tendency in Text-to-Speech Models . . . 63
5.3.2 Improving Tokenization . . . 64

References **65**

List of Figures

3.1 Count of pseudo-words in 4 splits. 18
3.2 Quadratic Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty. 29
3.3 Quadratic Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty. 30
3.4 Quadratic Regression (the last 3 splits) showing models' and humans' ways of feeling the difficulty. 31
3.5 Linear Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty in terms of distance to golden. 32
3.6 Linear Regression (the last 3 splits) showing models' and humans' ways of feeling the difficulty in terms of distance to golden. 33
3.7 Linear Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty in terms of EMD. 34
3.8 Linear Regression (the last 3 splits) showing models' and humans' ways of feeling the difficulty in terms of EMD. 35
3.9 Linear Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty in terms of bewilderment. 37
3.10 Linear Regression (the last 3 splits) showing models' and humans' ways of feeling the difficulty in terms of bewilderment. 38
3.11 Linear Regression (all 4 splits) showing models' and humans' ways of feeling the difficulty in terms of performance. 39
3.12 Linear Regression (the last 3 splits) showing models' and humans' ways of feeling the difficulty in terms of performance. 40

4.1 Vowel shifting in Monophthongs (pseudo-words). 53
4.2 Vowel shifting in Monophthongs (validation set). 54
4.3 Vowel shifting in Diphthongs (pseudo-words). 55
4.4 Vowel shifting in Diphthongs (validation set). 56

List of Tables

2.1 Transcription rules (vowels). 12
2.2 Transcription rules (consonants). 13

3.1 t-test on hit golden. 20
3.2 t-test on distance to golden. 21
3.3 Ranking on distance to golden. 22
3.4 Ranking on human-model distance. 23
3.5 t-test for model-human distance and inter-human distance. 24
3.6 Average distance to the closest human prediction. 25
3.7 Ranking on the distance to the closest human prediction. 25
3.8 The Earth Mover's Distance. 26
3.9 Ranking on the Earth Mover's Distance. 27

4.1 LLMs' mistakes regarding consonants (part 1). 44
4.2 LLMs' mistakes regarding consonants (part 2). 45
4.3 Validation results showing rigid mappings. 46
4.4 Validation results on silent consonants. 47
4.5 Validation results on semi-vowel j. 48
4.6 LLMs' mistakes regarding vowels (part 1). 49
4.7 LLMs' mistakes regarding vowels (part 2). 50
4.8 LLMs' mistakes regarding vowels (part 3). 51
4.9 Count of the occurrences of vertical shifting. 54
4.10 Comparing the gliding distance of diphthongs with t-test. 57
4.11 Count of cases where the length of prediction is influence by POI. 57
4.12 LLMs' mistakes regarding vowels intertwined with ɹ. 58
4.13 Count of cases where the LLM fails handle silent *e*. 59
4.14 Validation results about redundant schwa. 60
4.15 Validation results about failure to centralist a vowel. 60

Chapter 1
Introduction

1.1 Background and Our Contribution

Large language models (LLMs) are essentially pre-trained to perform language modelling, so it is not surprising to see them hold conversations. On top of that, LLMs seem to have learnt knowledge that is contained in their training data to the extent that, without being fine-tuned to do so, they can exhibit such knowledge in relevant tasks about math [28], [8], logic [37], [34], general knowledge [51], spatial reasoning [10] and more. They are also shown to possess linguistic knowledge about syntax and semantics (see Section 1.2.1).

Our study is interested in LLMs' knowledge about English grapheme-to-phoneme (G2P) rules and phonetic/phonological knowledge. English is notorious for its irregularity when it comes to G2P. That is, given the orthographic spelling (graphemic) of a word, the pronunciation (phonetic) can often be surprising (e.g. *gauge*, *indict*, *salmon*) and usually the reality is too complicated for any rule to be set out (e.g. the *eo* in *geography*, *theory*, *people*, *McLeod* all sound differently). Such is not the case in other Indo-European languages like Spanish and German, where the speaker can read the word based on its spelling without fail given the speaker has mastered the necessary rules.

However, this by no means suggests that native speakers of English do not have certain knowledge about G2P in their own language. In fact, when given a made-up "English" word with orthographic spelling, which we call a *pseudo-word*, native speakers are shown to be able to give a reasonable prediction of the pseudo-word's sound, and the predictions made by different human participants often agree with each other or converge to 1-2 possibilities [52], [39], [15]. The moral of the story is that native speakers are likely to possess knowledge about English G2P, more often than not implicitly and such knowledge cannot be described with a set of rules. When our native human participants unanimously decided that the vowel in the pseudo-word *nal* is [aː] rather than, say, [æ], they applied the knowledge, though they would likely report that they had been simply "guided by the gut feeling".

Whether LLMs have the same knowledge is questionable. In theory, modern day LLMs have word pronunciations included in their training data in the form of International Phonetic Alphabet (IPA) (these are available in Wikitionary, for example) and should be able to pick up relevant knowledge in the text modality. In fact, they have exhibited certain degree of ability on sound related tasks such as rhyming detection and syllable counting [54], [43]. Similarly, they can perform the G2P task as a sequence-to-sequence prediction task; [35] has shown that LLMs can give the pronunciation, in the form of ARPAbet[1], of a real word, and the accuracy is remarkable. We just cannot dismiss the

[1] ARPAbet and IPA are convertible except for minor caveats. See speech.cs.cmu.edu/cgi-bin/cmudict. See section 2.5 for the caveats.

concern that, on real English words, models could more or less memorise the correct pronunciation from their training data, whereas pseudo-words enable us to see if LLMs' knowledge learnt during the training are transferrable.

This type of knowledge differs from other types in two ways. 1) Unlike math, logic, common sense, facts, etc., pseudo-word G2P has no objective golden answer. We are measuring the alignment between LLMs and humans. In other words, we will examine how closely LLMs match human performance, which we call "human-likeness" in the rest of the thesis. 2) While we see nothing weird for LLMs to learn math, logic, common sense, facts, etc. by processing text, as humans do very much the same, native human speakers learn about G2P in a bi-modal fashion: they see the spelling and hear the sound. While it is discussed above that such process can happen in the modality of text, there also exist text-speech models that learn and work in the bi-modal fashion. We wonder if LLMs, much larger in size and stronger overall, fare any better than such text-speech models.

Therefore, we include 10 LLMs, one text-speech model (SpeechT5), and human participants in our study. At a high level, we address these questions:

Chapter 3: Quantitatively, how much human-likeness is there in LLMs' performance on the pseudo-word G2P task? How do LLMs compare with text-speech models on this front?

Chapter 4: Qualitatively, what are LLMs' tendencies when making the prediction that make them different from humans?

Correspondingly, our main findings are:

Chapter 3:
(1) LLMs' predictions are largely reasonable from a human perspective, but the distribution of model predictions is not phonetically close enough to human predictions, and models err on around 20% pseudo-words where human participants all give a unanimous golden prediction. This means some knowledge that human native speakers have is not encoded in LLMs. Also, model performance, in terms of hitting the golden prediction, is correlated with the human performance, indicating both groups feel the difficulty of a pseudo-word the same way; however, the correlation is not observed in terms of inter-human/run variation.
(2) While 3 LLMs generally considered outdated are trailing, the rest of the LLMs do not show significant difference between them; that is, state-of-the-art models like GPT5.4 are not outperforming small models like Llama3-8B-Instruct in terms of human-likeness. By comparison, SpeechT5, representing text-to-speech models in our study, prevails on all metrics. Put together, the results suggest there could be an inherent bottleneck for LLMs when in comes to learning G2P rules.

Chapter 4:
(3) We identified 3 common types of mistakes related to consonants and 6 related to vowels that LLMs tend to make on pseudo, and such are mostly validated using in-vocabulary words:
i. Rigid mapping to fricatives
ii. Failure to handle silent consonants
iii. Extra [j] mapped from *u*

iv. Frequent vertical shifting of monophthongs
v. Shortened glide of diphthongs
vi. Prediction length influenced by the length in the grapheme
vii. Failure to handle the silent *e*
viii. Extra [ə] in rounded vowels
ix. Failure to centralise unstressed vowels
Among them, (i) and (ix) both have to do with the models tendency to map a grapheme to a phoneme that is considered the most common. (ii), (vii), and (vi) all shows the influence of the grapheme in terms of the string length of the prediction; this can also be viewed as a form of rigidity. The rest phenomena are still mysterious and we try not to read too much into them.

Since this is, to our knowledge, the first study where LLMs, as well as the text-speech model, predict the sounds of pseudo-words, we believe the findings can

(1) reveal LLMs' actual knowledge about English G2P and pronunciation.
(2) shed light on future development of text-to-speech models and sound-to-IPA models (the latter helps field linguists with speech transcription).

1.2 Related Works

1.2.1 Linguistic Knowledge of LLMs

When it comes to syntax and semantics, behavourial tasks see satisfactory performance from modern day LLMs, as such tasks are what they are built for. As a result, most current studies focus on interpretability, using methods such as probing to decide if syntactic or semantic knowledge has been coded somewhere in the model weights. For example, [56] probed BERT and successfully located regions for recognizing the part-of-speech, constituents, dependencies, entity names, semantic roles, and co-reference. That is to say, the model has some standalone knowledge of these semantic and syntactic concepts.

Earlier, [47] not only managed to locate layers for morphological features, syntactic features, longer-range semantics respectively across a range of different architectures, but also shown that the constituent parsing tree can be predicted from certain layers in biLMs to some extent. A more comprehensive study is [12], where 10 tasks of different concerns, ranging from syntactic ones such as tree depth to morphological ones such as number and tense, were devised and correlation between the embedding and the performance were found in *some* cases. Similar studies along this line are [29], [30], [5], [32], and [9], to name a few important ones.

Though less studied than LLMs' syntactic and semantic knowledge, phonetic/phonological knowledge is also shown to be possessed by LLMs, but LLMs easily fall short of human native speakers on this front. Three tasks, namely G2P, syllable counting, and rhyming detection, are used in [54], and LLMs, ranging from Mistral-7B to GPT4 and Claude-3-Sonnet generally achieve an accuracy $\leq 60\%$ while humans never fall below 90%. There are similar findings in Greek [7] and Persian [48], and the fact that these languages are of lower-resource compared with English made the performance grimmer. It seems [35] is the only study that use a probe on phonetic/phonological tasks

and the result suggests that LLMs do capture relevant knowledge in deeper layers.

Our study, by comparison, is purely behavioural. Using pseudo-words - and also real words in validation experiments in Chapter 4 -, we not only show LLMs' human-likeness and non human-likeness in terms of G2P and phonetic/phonological knowledge, but also outline LLMs' tendencies. In future, we expect to explore the cognitive root of these tendencies with techniques such as probing. This way, we might have a better idea of how different sounds or G2P issues are captured by the model and what the possible reason is for certain tendencies.

1.2.2 Humans on Pseudo-word Tasks

In linguistics and cognitive science, pseudo-words have long been used to help us understand human language processing. The general underlying assumption is that processing pseudo-words is not much different from processing real words in the sense that knowledge and cognitive strategies for processing real words can carry over to pseudo-words. Therefore, success in processing pseudo-words is strong evidence that one is equipped with relevant knowledge. Also, since there will be more errors, latency, or inter-human disagreement when it comes to pseudo-words, researchers are able to peek through and better describe the (implicit) linguistic knowledge they are concerned with, which they cannot often do with real words.

The concerns of such studies fall in various subfields of linguistics. The pioneering study [6] claims that English speaking children have morphological knowledge since early years as they have no difficulty generating the plural form of pseudo-words. The claim is precisely based on the aforesaid assumption, which is also the assumption we hold today in our study, just that we are applying the assumption to LLMs and are concerned with phonology and phonetics. Other studies used pseudo-words to demonstrate human's knowledge of phonology [53] and [27], orthography [21], developmental phonology [22], lexicology [40]. Understandably, studies involving pseudo-words are focused on linguistic knowledge or facts below the syntactic level.

While long recognized as effective methodology, pseudo-words are seldom used in NLP. The only study we know is [17], where models like GPT-2 and BERT are shown to have the tendency to make the embedding of a pseudo-word closer to that of its definition than to the definitions of other pseudo-words - the "definition" is generated by human participants. This way, it suggests human-likeness in LLMs' semantic and morphological knowledge. In this regard, it is natural to move on to examine human-likeness in LLMs' phonetic/phonological and orthographic knowledge.

Chapter 2
Experiment Settings

We use the pseudo-word pronunciation task to see if LLMs, along with a few other models, have human-like knowledge on English pronunciation and G2P rules. The overall workflow is to 1) generate a set of pseudo-words, 2) get the predicted pronunciation from both the models and the human participants, and 3) compare the model and human predictions.

2.1 Generating Pseudo-Words

We want our pseudo-words to cover English G2P issues that are known to linguists as much as possible, while maintaining a reasonable workload for each participant to avoid deteriorated performance induced by a lengthy experiment. We finally made a list of 229 pseudo-words and each participant would get a random 48 out of them (see the following section). The list was made this way:

1. We included every known G2P issue mentioned in [23] and [57], be it an easy one like "*b* to [b]" or a tricky one like "*mb* to [m] in word final *mb*" (e.g. *comb, climb*). For each issue, we collected three relevant words from English vocabulary, one having a single syllable, one having two syllables, one having three or more. This was done with CMU Pronouncing Dictionary[1], which is downloadable and allows sub-string searches, and online word finder tools[2] and we double checked to make sure they are common English words rather than foreign words with short or zero history in English. If one issue did not have, say, a one-syllable example, we simply skipped and started with a two-syllable example. We call these words "base words" and the part of the word related to the issue in question is called the "part of interest" (POI). For example, for the issue "*mb* to [mb] in word final *mb*", a two-syllable example could be *succumb* while the POI is *mb*.

2. For each base word, we apply two operations. i) **preserving the structure**: If the POI is a consonant string (one or more consecutive consonant letters; we let C denote a constant string in this part only), all consonant strings in the word are preserved while all vowel strings (one or more consecutive vowel letters, including non-beginning *y*; we let V denote a vowel string) were randomly changed to other vowel strings. If the POI is V, we do the opposite. ii) **preserving the syllable** : All C or V except the CV combination that includes POI are changed to other C or V at random, C for C and V for V. That is, if the POI is V, then only the C immediately before it and itself are preserved. If the POI is C, then itself and the V following it are preserved. There are two additional rules:

[1] speech.cs.cmu.edu/cgi-bin/cmudict
[2] such as thefreedictionary.com/Word-Finder.htm

- The word-ending CVC is considered a syllable and entirely preserved when it comes to **preserving the syllable**.
- C should be further chopped up into two C's if it is not in a list of English consonant clusters (e.g. *dch* in *godchild* would be chopped up to *d* and *ch* while the final *ld* in this word was considered a whole - not because *ld* falls in our list, but simply because it is the end of the word). The list of consonant clusters is: *pl, bl, kl, gl, pr, br, tr, dr, cr, kr, gr, fr, tw, dw, fl, sl, cl, sc, wh, sw, th, sh, ch, sp, st, sk, sm, sn, sq, sp, ph, gn, kn, ps, cc, mm, nn, pp, tt, ff, gg, ll, rr, ss, thr, shr, sch, thw, sph, spl, spr, scl, str, scr.*

As an example, *happiness* (with the POI *pp*) can generate the pseudo-word *hoppeanass* at this step via **preserving the structure** and *strippilod* via **preserving the syllables**. Both would enable us to test models/humans on the "*pp* to [p] or [p^h]" mapping. For a base word, we collect up to 5 candidates via each operation, making a total of up to 10.

3. From the multiple candidates, we manually removed those that did not look orthographic before choosing 1-3 pseudo-words for each G2P issue (not for each base word). This made the first part of our pseudo-word list (184 words in total).

4. We also included a small set of exceptionally tricky words in terms of pronunciation that were collected from [23], [57], and [20]. These words feature G2P mappings that are unusual, if not unique, in English such as *museum*, where "*eu* to [iːə]" is the issue in question. For each word, we manually apply minimal change, usually by just substituting or adding 1-2 letters . For example, *taseum* stems from *museum* with *eu* being the POI. This made the second part of our pseudo-word list (45 words in total).

The reason why we apply minimal change to words in the second part is that we want to give the models and human participants a chance to spot the resemblance between the pseudo-word and the unique origin. If this happens, human participants tend to be affected by the origin when predicting the sound, using the cognitive strategy of analogy [24]; interestingly, we found through some preliminary exploration that state-of-the-art LLMs such as GPT5 and Gimini3 can also spot the resemblance, articulating such in their responses, while such reasoning is often decorative, the final prediction[3] having little to do with the sound of the origin word [4].

By contrast, as the majority of the list involves G2P rules known to linguists, we assume native speakers know them too, either explicitly or implicitly [19]. Some rules are unconditioned (such as "*sh* to [ʃ]") and can be tested with pseudo-words generated by either **preserving the structure** or **preserving the syllable**. Some rules are conditioned (such as "*c* to [s] before *i* or *e* but to [k] or [k^h] before *a*, *o*, or*u*"), and C is usually conditioned on the following V while V on the preceding C, which can be tested with pseudo-words generated by **preserving the syllable**.

[3]The word "final" can be misleading here. Actually, the prediction usually comes before the reasoning in LLMs' responses.

[4]For example, when given the pseudo-word *jestaurant* asked to "predict its pronunciation in IPA and give a brief reason", Gemini - the website - says "IPA: /ˈʤɛstərənt/ ... The word is clearly a portmanteau modeled after **restaurant** ... [W]hen a new word is created by swapping the onset of an existing word, the prosody (rhythm) and vowel reduction of the original word are typically preserved." Here, the model claims to use analogy, only to give an irrelevant prediction - the relevant one is /ˈʤɛstərɑnt/.

We should make it clear that, even for pseudo-words in the second part, we do not hold any "correct" prediction. For example, if the human predictions of *taseum* have little to do with its origin *museum*, we do not consider the participants as wrong. The model-human comparison is always based on actual human predictions rather than our prescriptive expectation.

Here is the full list of the 229 pseudo-words (the POI is bold)[5]:

gor**ps**	**ps**ymno	opla**mb**	nou**bt**	oske**bt**ed	fi**ck**	cho**ck**od
kneck	**kn**othy	**qu**il	le**qu**esh	**gh**est	a**gh**ach	**g**ourdien
di**gn**	calei**gn**	**ch**eme	ip**ch**arl	ke**ch**armi	sospa**tch**	leka**g**e
aba**g**ini	su**dg**e	olli**dg**e	wor**dr**abo	**dr**umotic	fy**th**	**th**easond
eidam	mou**th**erdam	be**sc**ich	o**x**i	e**x**eboha	oo**s**y	sa**zz**
ci**zz**y	ma**mn**	ci**ng**	jaso**ng**	re**ll**	hespi**ll**	houspa**ll**
ce**rr**y	ha**rr**ocona	**wr**eta	**wr**ithlob	**rh**ypo	**rh**one	nu**pp**ola
o**bb**	ri**bb**ir	la**gg**ato	she**ss**	cri**mm**er	o**nn**	nu**nn**y
ye**ll**ob	pod**e**	sizar**e**	ch**ee**be	wallp**e**te	th**e**mi	cath**e**krod
r**ea**t	c**ea**cob	deb**ea**mel	r**ie**pe	k**ei**t	s**ei**zume	m**ey**
kif**ey**	sp**i**	prest**i**li	n**i**k	c**i**ttos	m**y**mn	bet**y**
matol**y**mb	presh**y**	phaml**ie**	t**e**t	**e**llo	g**ea**d	sl**ea**sp
avy	gn**a**lf	j**a**tty	k**u**n	c**ou**sh	kl**oo**d	sc**oo**dy
s**a**ss	**a**gged	r**a**rsh	n**o**ge	**o**nooc	at**a**rs	pem**oir**
d**o**bd	d**o**ppy	f**o**cipay	g**au**ce	d**o**kt	amm**o**ck	n**a**l
awkbinn	all**ore**	f**ore**pinded	l**oo**r	r**oa**rpa	ig**oa**rd	m**ou**riph
s**u**t	t**o**lf	b**o**sal	f**oo**k	tosp**oo**n	wimper**oo**	m**o**lia
p**ou**mp	sn**ou**pie	zareb**ou**	r**u**ma	d**u**my	sh**ew**	embr**ew**
v**ue**	spat**ue**	eben**ue**	s**ui**vy	nen**oe**	b**ir**p	m**ir**ty
l**er**	g**ea**rn	**ea**rny	osk**ur**	n**u**rdes	s**u**rr	b**u**cry
dokn**i**tte	san**i**my	ek**e**n	s**a**ze	r**a**dy	d**u**gnore	ind**er**
sej**or**	osr**ai**n	ch**ey**	turb**ey**	h**i**cent	gl**y**	t**y**hoch
d**oe**sh	**ai**sda	th**ai**pomn	k**ie**	hopl**ie**d	c**oi**ston	d**oy**
gor**oy**	s**ou**t	sh**ou**tess	g**ow**	fr**ow**chy	ph**ee**l	**ee**gy
c**ea**r	ann**ea**r	v**e**de	hudc**ere**	r**ie**f	gh**are**	c**ai**r
th**eo**nely	suf**ai**r	wh**ea**r	mearing	cheoky	foor	moodist
tuce	oth**ure**	l**ou**rd	ba**ch**t	liccou**gh**	lecheip**t**	**pn**eudob
kos**c**le	vas**qu**ido	night**e**ouce	fol**d**ier	mi**gg**est	wis**t**ory	ban**d**ary
smi**ss**or	po**c**ean	dei**z**ure	frei**g**e	**g**enroi	te**ph**erd	jastaur**a**nd
b**ui**tal	beq**uay**	p**eo**tal	b**u**sitish	min**u**do	**e**ngtant	s**ai**p
sb**u**ry	g**eo**ffrant	br**ie**nd	fl**ai**t	g**ye**	kn**ow**lism	gr**oa**d
zr**ou**gh	g**yr**tle	sk**olo**nel	g**au**git	d**uo**y	tas**eu**m	vak**i**r
jewelim	t**uy**	sm**ei**r	l**ere**	resc**a**rce		

[5]One thesis committee member noticed that three items, namely *sass*, *vue*, and *shew*, are existing words and can be familiar to participants; *vue* is a common French word and *shew* is an archaic spelling of *show*. The vocabulary we checked the items against was the Word Lists split (size=757777) of the Natural Language Toolkit (NLTK) corpora (item 120 at https://www.nltk.org/nltk_data/); it turns out that we should have checked against multiple vocabularies. Meanwhile, we believe the inclusion of the three items has not skewed our main research findings in any appreciable way.

2.2 Human Participants

We recruited online human participants on Prolific[6], where users provide their personal information upon registration. We set the filter such that "Canadian or US born" users "aged between 18 and 38" whose "primary and first language is English" are made eligible. We recruited 20 participants, who all speak typical North American English.

Each participant was told the purpose of the experiment beforehand and encouraged to give a "reasonable prediction" for each pseudo-word. They were allowed to ponder for a few seconds or change their prediction, by recording again, before moving on to the next trial. On each trial, they saw one pseudo-word appearing at the centre of the screen and their predicted pronunciation was recorded through the microphone. They were instructed to say each word in a "down tone" (falling intonation) and avoid an "up tone" (rising intonation); example audios of both tones were played to them. The purpose is to ensure the natural stress pattern. If a multi-syllable word is pronounced with rising intonation, the final syllable can feature a higher pitch and extra muscular energy, usually leading to difficulty in deciding which syllable is meant to be stressed.

Each participant went through 60 items, among them 12 were existing English words. They serve two purposes: 1) Attention check and 2) Eliciting their pronunciation of the primary allophones of English, which would be referred to when we transcribed their responses. The 12 words remained the same across participants:

please, great, bath, shop, slip, pants, rat, pinch, such, mouth, neck, no

The participants were instructed to simply say the word whenever they decided it was a "real English word". The rest 48 items were randomly selected from the 229-strong list of pseudo-words.

2.3 Models

We have 4 groups of models.

Group 1 is 5 small LLMs: Llama-3-8B-Instruct [25], Mistral-7B-Instruct-v0.3 [33], Qwen-2.5-7B-Instruct [59], GPT-4o-mini (model size undisclosed; estimated to be 8-12B)[31], and Gemini-2.5-Flash (model size undisclosed; estimated to be 20-30B) [11].

Group 2 is 2 flagship LLMs from the GPT and Gemini families at the time of writing: GPT5.4 [44] and Gemini-3.1-pro [18].

Group 3 is 3 older small LLMs generally considered weaker: Falcon-7B-Instruct [1], MPT-7B-Instruct [55], Dolly-v2-12B [13]. They are shown to be weaker than Group 1, not to mention Group 2, on tasks such as question answering [16] and [50], math reasoning [33], common sense knowledge

[6] app.prolific.com

[41], and instruction following [38].[7]

In the rest of the paper, the LLMs mentioned so far are referred to as Llama3, Mistral, Qwen2, GPT4, Gemini3, GPT5, and Gemini4 respectively. Note that the final 4 models are more or less monotonically ascending in terms of capacity, forming a complete scale that allows us to comprehensively see LLMs' performance on our G2P task.

Group 4 contains only one text-to-speech (TTS) model: SpeechT5 (144M) [2]. It is small but dedicated to six bi-model (audio and text) tasks, text-to-speech being one of them - G2P is essentially a text-to-speech task for the model. It uses a shared embedding space for audio and text. Its pretraining not only includes tasks in both modalities, but also encourages the model to rely on the modality-invariant embeddings as much as possible. [8]

Group 5 contains 3 models/algorithms more of a baseline nature: an BiLSTM G2P model we trained by ourselves, a similar "g2pE" model [46] with the Gate Recurrent Unit (GRU) architecture dedicated to English G2P tasks, and a probabilistic baseline algorithm that drew on the co-occurrence frequency of graphemes and phonemes.

The three are referred to as myG2P, kbG2p, and baseline in the rest of the thesis.

myG2P is a BiLSTM model (with attention) with Latent Dimension= 256 and Embedding Dimension= 128. It is trained on a sequence-to-sequence prediction task using the CMU Pronouncing Dictionary[9]. We exclude all words with ≤ 3 letters because they are more often than not abbreviations such as AAA ("T-R-IH-P-AH-L-EY"), which are obviously unclean data on the G2P task. During training, the input is the word and the expected output is the pronunciation of the word in ARPAbet (not IPA, as the CMU Pronouncing Dictionary uses ARPAbet to encode the sound). We trained for up to 200 epochs with dropout rate= 0.3. (With tolerance= 0.001 and patience= 15 epochs, it actually stopped at Epoch 38.)

The baseline algorithm is also based on the CMU Pronouncing Dictionary. First, two "_" symbols denoting string boundaries are added to the front and end of both the word and the pronunciation string. Next, each word as well as each pronunciation string is chopped up into the form of "_(C)V, CV, ..., CV(C)_". We call each "CV" a chunk. For example, _*melody*_ has 3 chunks, namely _*me*, *lo*, *dy*_; its sound in ARPAbet _M-EH-L-AH-D-IY_ also has 3 chunks, namely _M-EH, L-AH, D-IY_. After doing this to all words with ≥ 4 letters, we have a list of all letter chunks and a list of all sound chunks. Then we count the times of co-occurrence for each letter chunk and each sound chunk. If both chunks appear in a word-sound pair, it counts as co-occurrence. For example, in _*me-lo-dy*_ - "_M-EH, L-AH, D-IY_", _*me* and "D-IY_" has the count+1 even though they are not matching parts. This way, we get a co-occurrence matrix, with each row being a letter chunk and each column being a sound chunk. Now for each letter chunk (a row), we decide its final prediction this way: First, the top 10 entries in the row are selected. Next, for each of the 10 entries, we

[7] We also tested Deepseek-coder-7B-Instruct-v1.5 [26] but the model even struggled in understanding the task and about half the responses were without a prediction of the pronunciation, so we discarded its results.

[8] In fact, we also tested a larger MetaMMS (330M) [60], but the responses on the current task were barely recognizable. One possible explanation is that SpeechT5 was pre-trained mainly on English data, but MetaMMS was not. We saw no point including MetaMMS in our comparison.

[9] speech.cs.cmu.edu/cgi-bin/cmudict

compute:

$$\frac{\textbf{entry}}{\textbf{sum of the column of the entry}}$$

If this value goes high, it means the sound chunk not only co-occur with the letter chunk plenty of times, the reason is actually the sound chunk is characteristic of the letter chunk, rather than, say, the sound chunk is overall common in English. The entry with the highest value is selected as the final prediction for the letter chunk. This way, we have an answer key, which allows the algorithm to predict the sound of a word. (Of course, the algorithm will need to chop up the word first and assemble the predicted sound chunks in the end.)

2.4 Model Inference

For all LLMs (groups 1-3), all 229 pseudo-words are plugged into a few-shot prompt one by one [10]:

Given a pseudo-word, predict its pronunciation in IPA assuming it is North American English.

input: jag
output: /ʤæg/

input: mig
output: /mɪg/

input: plud
output: /plʌd/

input: veel
output: /viːl/

input: [pseudo-word]
output:

Considering that prompts could affect the response and that we wanted to see a distribution of responses across any given model just like we did with the human participants, we ran each LLM 4 times with 4 distinct sets of few-shot examples (that is, the templates are the same but the example inputs/outputs are different).

For SpeechT5, we also run it 4 times, each time using a different speaker embedding. It is part of the model functionality that one can select the "speaker". Not only does the timbre change, the

[10] In our prompts, each sample prediction is enclosed in a pair of slashes ("/viːl/"). Strictly speaking, the prediction is at the level of phones rather than the that of phonemes and, by the IPA transcription convention, should be enclosed in square brackets ("[viːl]"). However, since a lot of online dictionaries and resources (including Wiktionary, which is bound to be inside most LLMs' pre-training data) use a pair of slashes for the pronunciation (see https://www.oxfordlearnersdictionaries.com/ and https://www.ldoceonline.com/, for example), we keep using the symbol in our prompts to enable the model to leverage what it has seen during pre-training.

prediction can also vary as the model's workflow is purely text-to-wave. The input is the pseudo-word alone and the output is an audio file. Though each run is deterministic, changing the speaker could have an effect on the prediction.

Put simply, for each LLM/SpeechT5, we got 4 predictions on each pseudo-word.

For the two G2P models and the baseline algorithm, the settings are all deterministic and the input was the pseudo-word alone. We get 1 prediction on each pseudo-word.

2.5 Transcription

For all LLMs, an output contained an IPA string somewhere in the response and, fortunately, all these strings were well structured so that we could always figure out the segment inside the string that was corresponding to the POI. In the example of *taseum* with the POI *eu*, if the output look like [tɛˈsɪjuːm], we decide that the pronunciation of *eu* is predicted as [ɪjuː] and only this segment is logged. In other words, the model was not penalised if other parts of the word were poorly predicted.

For the two G2P models and the baseline algorithm, the output is an ARPAbet string. We then convert ARPAbet to IPA, but there are two caveats: 1) both /ʌ/ and /ə/ are denoted by AH in ARPAbet, and 2) both /ɜɹ/ and /əɹ/ are denoted by ER in ARPAbet. In other words, ARPAbet does not include a dedicated symbol for /ə/. As it would be unfair to simply assign the more reasonable possibility during transcription, we reluctantly chose to log these as /ə/ and /əɹ/. The effect on training, however, can not be undone by any means. Therefore, we can say the two G2P models and the baseline are somewhat disadvantaged.

For SpeechT5 and human participants, an output is an audio file. We transcribed each file and also logged the segment that matched the POI. To ensure that the comparison was fair and the data were suitable for follow-up computation using PanPhon (see the coming section), we stuck to a set of transcription rules, which are detailed out in Table 2.1 and 2.2 .

Remark 1: At the end of a pseudo-word like *phamilie*, if the prediction is [ɪ] or [iː], they are kept as they are and penalised (in the sense that most humans and models would predict [i] here).
Remark 2: PanPhon treats all long vowels as one vector.
Remark 3: [e] is logged as [ɛ]. For the vowel in the word *bed*, it is usually [ɛ] in the North American English and a lower [e] in the British English ([ɛ] "-tense" versus [e] "+tense"). We decide not to penalise the model for predicting [e] when the humans' decision is [ɛ]. But cf. remarks 8 and 10.
Remark 4: In theory, sounds like [ɑ̃] or [ɔ̃] in French words are considered to have become [ɑː] when borrowed by English. We keep these nasal vowels if they do occur, although there is no such case in the current experiment.
Remark 5: During audio transcription, the [ɑː] and [ɔː] sounds in pseudo-words like *doppy* and *dokt* are logged as [ɒ] since they are allophones of the same phoneme - this should be confirmed using elicit words for every participant. The same applies to logging LLMs' predictions.
Remark 6: Failure in centralisation is penalised as it is considered part of the phonetic/phonological knowledge. For example, whether or not humans' prediction for the pseudo-word *sanimy* is [ə], [ɪ]

Loose IPA	Logged as	Remarks
i	i	1
iː	iː	2
ɪ	ɪ	
e	ɛ	3
æ	æ	
ʌ	ʌ	
ɑː	ɑː	4
ɔː	ɔː	
ɒ	ɒ	5
ʊ	ʊ	
uː	uː	
ɜː	ɜː	
ə	ə	6
eɪ	eɪ	7, 8
aɪ	aɪ	
ɔɪ	ɔɪ	
əʊ	oʊ	9
ɪə	ɪə	
eə	ɛə	10
ʊə	ʊə	

Table 2.1: Transcription rules for vowels. The rule of thumb: LLMs are only penalised for a "wrong" judgement of the phoneme. But since they follow the loose IPA convention (the one most dictionaries use), which is reasonable, they should not be penalised for failing to specify the allophone if the phoneme is correctly predicted. We manually change their prediction to the "correct" allophone. We only do this to make the vectors as precise as possible in the follow-up analysis.

as a model prediction is logged as it is. Similarly, in pseudo-words like *oskur*, the prediction [ɜːɹ] is considered different from [əɹ]. It is a piece of knowledge that in most cases a long vowel is unstressed. Cf. 1 and 19.

Remark 7: All diphthongs count as two separate monophthongs in PanPhon.

Remark 8: [eɪ] is not changed to [ɛɪ] as the former is considered the "standard" in North American English. We simply log it as [eɪ]. Cf. 3 and 10.

Remark 9: The British [əʊ] and American [oʊ] are both logged as [oʊ]. They are two allophones of a phoneme and the difference should not be penalised.

Remark 10: [eə] is changed to [ɛə]. Cf. 3 and 8.

Remark 11: During transcribing, aspiration is explicitly marked whenever it occurs. Nearly all model outputs do not include aspiration information but this should not be penalised as dictionaries do not usually mark it, too. Therefore, we manually mark the aspiration where it "should" occur - not after [s]. Cf. 12 and 13.

Remark 12: In such pseudo-words as *nighteouce*, participants can predict the sounds as [ɾ]. If this is the case, it is transcribed as [t^h]. The [ɾ] sound is hardly reflected in dictionaries and if the

Loose IPA	**Logged as**	**Remarks**
p	p, p^h	11
b	b	
t	t, t^h, ɾ	12
d	d	
ʔ	ʔ	13
k	k, k^h	
ɡ	ɡ	
ʧ	ʧ	14
ʤ	ʤ	
ʃ	ʃ	
ʒ	ʒ	
tr	ʧɹ	
dr	ʤɹ	15
f	f	
v	v	
θ	θ	
ð	ð	
s	s	
z	z	
h	h	
m	m	
n	n	
ŋ	ŋ	
l	l, l̥, ɬ	16
r	ɹ	17, 18
j	j	
w	w	

Table 2.2: Transcription rules for consonants. The rule of thumb: LLMs are only penalised for a wrong judgement about the phoneme. But since they follow the loose IPA convention (the one most dictionaries use), which is reasonable, they should not be penalised for failing to specify the allophone if the phoneme is correctly predicted. We manually change their prediction to the correct allophone. We only do this to make the vectors as precise as possible in the follow-up analysis.

models predict [t], which they mostly do, we should not penalise them and log that as [t^{h}]. Strictly speaking, [t^{h}] and [ɾ] are not interchangeable in many contexts, but, as said, [ɾ] is not part of the loose IPA and model are not to blame. Cf. 11.
Remark 13: In fact, in words like *button*, a plosive can, and perhaps should, become [ʔ]. Thus, if a model predicts [ʔ] somewhere, we log it as it is and it can be penalised. However, this piece of knowledge is not involved in the current experiment. Cf. 12.
Remark 14: PanPhon includes t͡ʃ, as well as d͡ʒ, as one vector.
Remark 15: In words like *trip* and *dress*, dictionaries use tr and dr - each is treated as one dedicated symbol. However, PanPhon use d͡ʒɹ and d͡ʒr (tɹ is broken down to t + ɹ and the same for dɹ). We use d͡ʒɹ and d͡ʒr accordingly.
Remark 16: PanPhon separates [l] (*lab*, -continuant), [ɬ] (*bell*, 0 continuant) and [l̩] (*Sybil*, +syllabic); the last type is missing from the current experiment. This is hardly reflected in dictionaries and LLMs rarely predict other than [l]. We manually separate the three during transcribing/logging according to the location of the sound in concern.
Remark 17: The correct IPA symbol should be ɹ while most dictionaries use r for convenience. We change r to ɹ accordingly during transcribing.
Remark 18: Models sometimes give ɚ, əˤ, etc. and these are all changed to əɹ. Actually, there is tiny difference in how one speaker produces a rhotic vowel sound but such by no means imply the difference in pronunciation or orthography knowledge. However, in such pseudo-words like *mearing*, the sound of interest is transcribed as [ɪɪ]; if the LLM predicts [ɪəɹ], it is recorded as it is and penalised. That is simply because [ɪɪ] hardly involves centralisation. Cf. 6.

2.6 Computing the Editing Distance

Now we had recorded predictions of all POIs from all models and human participants, each as a short IPA string, we wanted to see to what extent each models' prediction is aligned with humans'. We noticed that "right or wrong" might be oversimplified. For example, for a given POI, if all human participants agree on sound [k] while two models predict [lj] and [g] respectively, they were wrong to different extents. We would say [g] is not to far off from humans' golden prediction and it would be ideal if this prediction is more lightly penalised than [lj] is. Therefore, we used PanPhon [42] [11] to compute the distance between two IPA strings.

PanPhon turns every phone in IPA to a feature vector of 24 dimensions. Each feature is coded as "+", "−", or "0". For example, on feature "high", [i] takes value "+" and [ə] takes "−" (since a mid vowel is "not high"), while [p] takes "0" as it is not a vowel, making the "high-mid-low" classification irrelevant. Note that a multi-phone string such as [dɑːt] is turned in to a list of multiple vectors. We used PanPhon's "weighted feature edit distance" function to find the distance between two IPA *strings* (not phones). Note that the transcription rules set out in the last section ensure that our logged IPA strings are PanPhon-ready.

Also note that in some cases the model or the participant could predict a POI to be silent, such as the ending *e* in pseudo-word *pode*. Unfortunately, PanPhon's phone inventory does not contain a "NULL phone". We resolve the issue by coding the silent phone as a 24-dimensional zero vector

[11] github.com/dmort27/PanPhon

and finding the total Euclidean distance between the string's phones and the origin. To be specific: for a non-silent IPA string $p_1, p_2, \cdots, p_n$, where p_i is a convertible phone, we convert it to a list of vectors $v_1, v_2, \cdots, v_n$. Then we retrieve PanPhon's weight vector w. According to how PanPhon computes the weighted edit distance, the total distance in this case should be

$$\sum_{i=1}^{n} \|v_i \odot w\|_2 + n$$

where $\odot$ means element-wise multiplication. The $+n$ at the end is the penalty for insertion (each insertion/deletion is penalised by $+1$ in PanPhon).

This way, we can compute the distance between any two predictions, which enables us to compare the performance of models and human in the next chapter.

Chapter 3
Analysing the Performance

In this chapter, we quantitatively examine how human-like model predictions are. "Human-likeness" can be understood in two ways:

1. Assume there is some golden prediction from the human perspective, and compare how close models and humans are to the golden prediction respectively. In our study, we use the majority human prediction (one that $\geq 50\%$ humans make) as the golden prediction. On a pseudo-word, if all human predictions are split down the middle, we have two golden predictions. For a pseudo-word where the average inter-human distance is close or equal to 0, the golden prediction(s) can give us a clear idea of humans' knowledge; when the inter-human distance goes higher, meaning the pseudo-word is relatively difficult, the golden prediction represents a popular, therefore reasonable, understanding of G2P among humans, but we should refrain from claiming that anything away from it is a non human-like prediction. In theory, it is possible for a model to beat humans on this front - if the non-golden human predictions are very far-flung while the model's prediction is close to the golden one - just like machines could resemble a human subset that are more sensible, intelligent, etc. than the average human being.

2. Dismiss the idea of golden prediction, and examine how close model predictions are to human predictions. It is meaningless comparing the models and humans on this front as humans are infinitely close to themselves; this type of metric, however, can be used for inter-model comparisons.

To measure the first type of human-likeness, we use two metrics: 1) the average number of times a model gives a golden prediction (if there are two golden predictions, giving either one will count) and 2) the average distance between a model prediction and the *closest* golden prediction. The second is a finer-grained version of the first.

To measure the second type of human-likeness, we use three metrics: 3) the average distance between a model prediction and a human prediction, 4) the average distance between a model prediction and the *closest* human prediction, and 5) the Wasserstein distance (Earth Mover's Distance) between the model predictions and the human predictions. Among them, 3) is essentially a measurement of how close the model is to an "averaged" human (see discussion in the coming section); 4) is a better measurement of whether the model can, to say the least, give some prediction that is conceivable to humans; 5) is the best indicator of how similar the two distributions represented by model predictions and human predictions are.

With these metrics at hand, we raise and address these research questions:

Question 1. How human-like are LLMs in the first sense?
Question 2. How human-like are LLMs in the second sense?

Question 3. Do LLMs and humans feel the difficulty of a pseudo-word the same way? (That is, do LLMs tend to predict worse or less confidently upon pseudo-words that see larger inter-human distances?)

The following section outlines the methods in detail and the next three sections address the three questions.

3.1 Metrics and Methods

We first need to decide the golden prediction(s) for each pseudo-word. If a prediction is made by ≥ 2 humans and represents $\geq 50\%$ of all human participants who were presented with this pseudo-word, this prediction is a golden prediction of this pseudo-word. A pseudo-word can therefore have no golden prediction when there is considerable inter-human disagreement.

Inter-human disagreement is measured by the average **inter-human distance**. For a given pseudo-word that receives >= 2 human responses, we compute pair-wise distances between these human predictions and then compute the mean. If and only if all humans decide on a single prediction, the mean is $= 0$. There are 216 pseudo-words that received ≥ 2 human responses[1] and 207 of them have one or more golden prediction. In our study, we use this measurement to decide how "difficult" a pseudo-word is and divide all the pseudo-words into four splits: average **inter-human distance** $= 0$, $\in (0, 2]$, $\in (2, 5]$, > 5. These are detailed out in Figure 3.1. For context, the distance between [iː] and [ɒ] - two most distant vowel phones - is 2.5, the distance between [p] and [h] - two most distant consonant phones - is 4.5, and the distance between sound strings [iːp] and [hɒ] is 15.75; the greatest average **inter-human distance** in our data is 8.0.

The five aforesaid metrics are computed this way:

1) The average number of times a model gives a golden prediction (referred to as **hit golden** in the chapter): for every model (or baseline), with either 4 runs or 1 run, there are 4 or 1 predictions on a given pseudo-word that has some golden prediction. If a prediction is golden, it gets score= 1; otherwise, it gets 0. This way, there exists an average score $\in [0, 1]$ across all runs. For all 207 pseudo-words that have some golden prediction, each model returns a 207-item array. We do exactly the same for humans, getting another 207-item array. Later, we can perform the t-test for any model array and the human array.

2) The average distance of a model's predictions and the *closest* golden prediction (referred to as **distance to golden**): for every model prediction, we find the distance between it and its closest golden prediction. On a given pseudo-word, there exists an average distance across all runs. This gives a 207-item array. Humans give us another 207-item array. These arrays are again t-test-ready.

[1] We believe if a pseudo-word has only 1 human response, it is unsafe to carry out analysis based on this sole prediction. If we had had more human participants, we would have set the threshold higher. As it is, 2 is the best we can do to avoid discarding too many data.

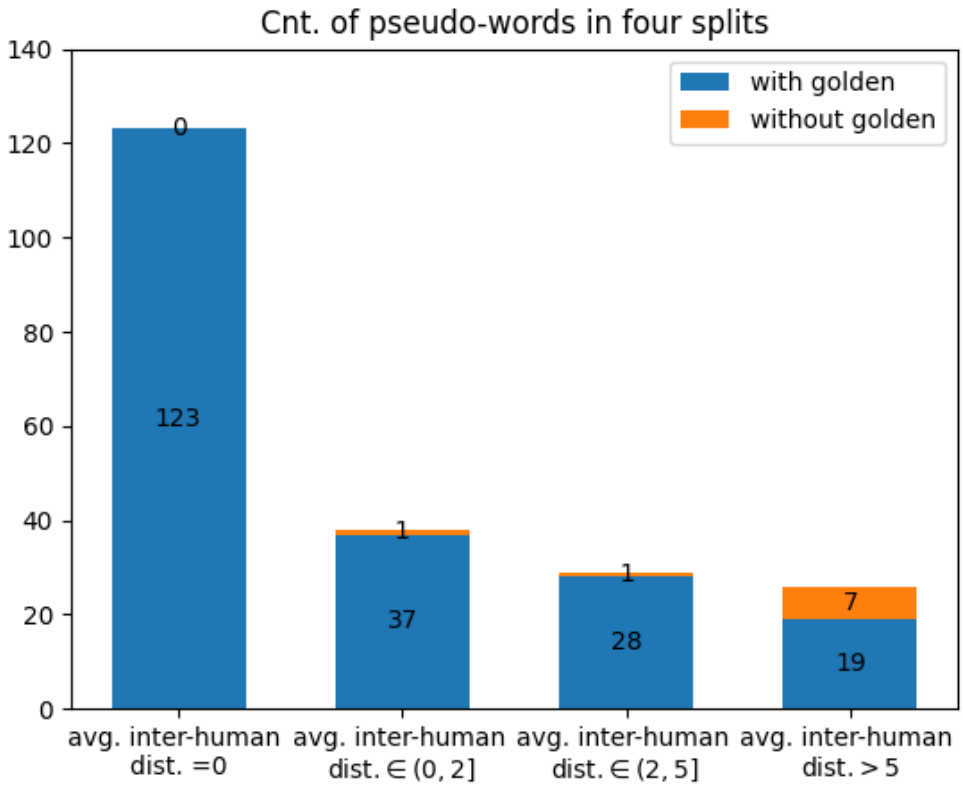


Figure 3.1: We use the average inter-human distance to measure a pseudo-word's difficulty. If a human prediction has ≥ 2 and $\geq 50\%$ votes, it counts as a golden prediction of the pseudo-word. Out of 216 pseudo-words with ≥ 2 human predictions, 207 have one or more golden prediction.

3) The average distance between a model prediction and a human prediction (referred to as **model-human distance**): On a given pseudo-word with ≥ 2 human responses, there is a distance for every pair of "model prediction vs human prediction". Average all these distances for this pseudo-word and we get a 216-item array.

Now we should ask: If we have two models, each giving us a 216-item array, and we perform a t-test for these two arrays, in what sense does the result tell whether one model is more human-like than the other? For any given pseudo-word, if all the human predictions are highly convergent (with one cluster), the measure is not very different from the **distance to golden** and can be interpreted this way. If human predictions are multi-modal (≥ 2 clusters) or even more scattered, than a group of model predictions where each is close to one of the human predictions can have the higher average **model-human distance** than a group of model predictions that are concentrated somewhere inside the "convex hull" of all human predictions[2]. But the first group of model predictions are obviously more human-like. Therefore, this measurement actually tells how good the model is at picking somewhere that is no too far from all human predictions - more intuitively, the tendency that a model resembles the average of all human beings, even when such an abstract human being does not exist.

4) The average distance between a model prediction and the *closest* human prediction (referred to as **closest model-human distance**): On a given pseudo-word, every model prediction has its closest human prediction - it could be an unpopular human prediction - and gives a distance. We compute the average distance across all runs and get a 216-term array. In contrast to 3), this measurement indicates how the model predictions are close to some real, rather than abstract, human predictions. The shortcoming is that the distributional similarity is not reflected.

5) The Wasserstein distance (Earth Mover's Distance) between the model predictions and the human predictions (referred to as **EMD**): This is the most direct measurement that serves the second meaning of "human-likeness". On each pseudo-word, there are 4 or 1 model predictions and ≥ 2 human predictions. We assume these two groups of predictions are sampled i.i.d. from two distributions and the similarity of the two distributions is reflected by the EMD between these two samples. In total, we get an array of 216-term array. (Wasserstein distance measures the minimal total distance required to move the probability mass of one distribution to another distribution, reflecting the similarity of the two distributions in a metric space [45].)

To recap, 1) and 2) measures how the model resembles the most sensible human being, 3) measures how the model resembles the abstract averaged human being, 4) measures how good the model is at giving some reasonable prediction, 5) measures how the model resembles the overall human beings distribution-wise.

[2]This is an intuitive and highly informal way of thinking about it, as there is no vector space of any kind. We can only define a metric space by computing the distance for any two predictions.

3.2 Human-likeness in Terms of Performance

Metrics 1) **hit golden** and 2) **distance to golden** are discussed here. As discussed above, we perform the two-tailed t-test to see if models surpass humans. The results are shown in Tables 3.1 and 3.2.

Model	**avg. inter-human dist.** $= 0$		$\in (0, 2]$		$\in (2, 5]$		> 5		**overall**	
	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p
Llama3	-0.260	**0.000**	-0.194	**0.020**	-0.091	0.289	-0.127	0.251	-0.213	**0.000**
Mistral	-0.331	**0.000**	-0.207	**0.016**	-0.108	0.181	-0.245	**0.039**	-0.271	**0.000**
Qwen2	-0.348	**0.000**	-0.268	**0.001**	-0.153	0.062	-0.284	**0.004**	-0.301	**0.000**
Gemini2	-0.161	**0.000**	-0.059	0.435	***0.061***	0.423	-0.113	0.229	-0.108	**0.000**
GPT4	-0.266	**0.000**	-0.194	**0.010**	-0.153	0.067	-0.206	**0.017**	-0.232	**0.000**
MPT	-0.461	**0.000**	-0.403	**0.000**	-0.296	**0.004**	-0.627	**0.000**	-0.444	**0.000**
Dolly	-0.608	**0.000**	-0.498	**0.000**	-0.341	**0.000**	-0.640	**0.000**	-0.555	**0.000**
Falcon	-0.604	**0.000**	-0.566	**0.000**	-0.305	**0.001**	-0.574	**0.000**	-0.554	**0.000**
Gemini3	-0.189	**0.000**	-0.099	0.229	-0.010	0.898	***0.005***	0.959	-0.131	**0.000**
GPT5	-0.197	**0.000**	-0.180	**0.032**	-0.055	0.484	-0.074	0.434	-0.164	**0.000**
SpeechT5	-0.124	**0.000**	-0.066	0.356	-0.082	0.340	-0.166	0.118	-0.112	**0.000**
kbG2P	-0.309	**0.000**	-0.113	0.172	-0.082	0.377	-0.087	0.479	-0.223	**0.000**
myG2P	-0.276	**0.000**	-0.005	0.951	-0.224	**0.024**	-0.245	0.054	-0.218	**0.000**
baseline	-0.520	**0.000**	-0.383	**0.000**	-0.367	**0.001**	-0.350	**0.005**	-0.459	**0.000**

Table 3.1: Two-tailed t-test that compares the average number of times of giving a golden prediction between models and humans (i.e. model − human). The model mean are human mean are both $\in [0, 1]$, so the difference is $\in [-1, 1]$ and the greater it is, the less inferior the model performance is.

Overall, LLMs are well beaten in Table 3.1 (**hit golden**), mainly because even in the easiest split, which is also the largest one, the golden unanimous prediction among human participants still eludes Gemini2, Gemini3, and GPT5 for 15-20% of the time for, no to mention other LLMs. However, looking at the three harder splits, we find that Gemini2 and Gemini 3 are not worse than humans, and Llama3 and GPT5 also catch up when it comes to the two hardest splits. Interestingly, models as a whole group are in their prime on the split of $\in (2, 5]$, not the hardest split.

When we use the finer-grained measurement of **distance to golden** in Table 3.2, we see quite the same picture. Most LLMs are not inferior to humans on the final two splits and, for once, Gemini3 defeated humans on the hardest split. From the two tables, the high-level message is:

Model	avg. inter-human dist. $= 0$		$\in (0, 2]$		$\in (2, 5]$		> 5		**overall**	
	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p
Llama3	1.336	**0.000**	1.192	**0.007**	0.096	0.873	0.500	0.606	1.066	**0.000**
Mistral	1.448	**0.000**	1.428	**0.003**	0.631	0.382	0.427	0.572	1.240	**0.000**
Qwen2	1.530	**0.000**	1.060	**0.008**	0.045	0.915	1.327	0.084	1.226	**0.000**
Gemini2	0.745	**0.000**	0.703	**0.023**	***-0.284***	0.598	0.095	0.872	0.538	**0.000**
GPT4	1.228	**0.000**	0.935	**0.008**	1.363	**0.043**	0.161	0.715	1.096	**0.000**
MPT	2.253	**0.000**	2.128	**0.000**	2.428	**0.009**	2.306	**0.014**	2.259	**0.000**
Dolly	2.851	**0.000**	2.730	**0.000**	0.795	0.145	1.812	**0.026**	2.456	**0.000**
Falcon	2.849	**0.000**	2.696	**0.000**	1.760	**0.021**	2.566	**0.001**	2.648	**0.000**
Gemini3	0.823	**0.000**	1.090	**0.025**	0.062	0.912	***-1.092***	**0.030**	0.592	**0.001**
GPT5	0.993	**0.000**	1.321	**0.005**	0.218	0.688	0.082	0.898	0.863	**0.000**
SpeechT5	0.343	**0.002**	0.418	0.115	0.672	0.265	0.437	0.596	0.410	**0.002**
kbG2P	1.206	**0.000**	0.674	0.087	***-0.434***	0.412	***-0.128***	0.859	0.767	**0.000**
myG2P	1.008	**0.000**	0.662	0.081	0.906	0.200	0.220	0.777	0.860	**0.000**
baseline	3.513	**0.000**	3.711	**0.000**	3.202	**0.004**	1.613	**0.026**	3.332	**0.000**

Table 3.2: Two-tailed t-test that compares the average distance to the closest golden prediction between models and humans (i.e. model − human). The model mean or human mean is ≥ 0, so the difference is in theory unbounded and the lower it is, the less inferior the model performance is.

Conclusion 1. When pseudo-words are hard, both LLMs and humans are bewildered and some humans can miss the golden prediction. But we can say LLMs are unlike humans in the sense that when pseudo-words are easy enough for humans to pronounce some obviously reasonable prediction in unison, LLMs still decide on something else.

From both tables, probably we can tell some LLMs (like Gemini2 and Gemini3) are better while some (like MPT, Dolly, and, Falcon) are not as good. To examine this, we perform the t-test between every pair of models and rank them using the Khan's algorithm[3]. We do this only for the **distance to golden** measurement and the result is in Table 3.3. The high-level conclusion are as follows:

Conclusion 2. On the easiest words, LLMs are not as accurate as SpeechT5, which represents the text-to-speech models.
Conclusion 3. Gemini2, Gemini3, GPT4, GPT5, and Qwen2 more or less make up the "first-tier", follwed by Llama3 and Mistral, while MPT, Dolly, and Falcon struggle the most, in many cases not even beating the baseline.
Conclusion 4. Though Gemini3 is generally bigger and stronger than Gemini2, it fails to prevail on the pseudo-word pronunciation (or G2P) task.
Conclusion 5. When compared with both simpler G2P models, LLMs do not fare any better.

[3]In theory, t-test is not transitive; the "winner-loser" relations can form a cycle. Fortunately, this does not happen in our study, so ranking is possible.

Split	Full Ranking
avg. inter-human dist. $= 0$	SpeechT5 $>$ Gemini2 = Gemini3 = GPT5 = kbG2P = myG2P $>$ GPT4 = Llama3 = Mistral = Qwen2 $>$ Dolly = Falcon = MPT $>$ baseline
$\in (0, 2]$	Gemini2 = Gemini3 = GPT4 = kbG2P = myG2P = Qwen2 = SpeechT5 $>$ GPT5 = Llama3 = Mistral = MPT $>$ baseline = Dolly = Falcon
$\in (2, 5]$	Gemini2 = Gemini3 = GPT5 = kbG2P = Llama3 = Mistral = myG2P = Qwen2 = SpeechT5 $>$ Dolly = Falcon = GPT4 = MPT $>$ baseline
> 5	Gemini3 = kbG2P = Llama3 = myG2P $>$ Gemini2 = GPT4 = GPT5 = Mistral = Qwen2 = SpeechT5 $>$ baseline = Dolly = Falcon = MPT
overall	Gemini2 = Gemini3 = kbG2P = SpeechT5 $>$ GPT4 = GPT5 = Llama3 = Mistral = myG2P = Qwen2 $>$ Dolly = Falcon = MPT $>$ baseline

Table 3.3: Ranking the model performance by the average distance to the closest golden prediction. Pair-wise comparisons are based on the t-test and Kahn's algorithm is applied to give the full ranking. The smaller the distance, the higher the ranking is. Models in the same tier (connected by =) are not significantly different on the t-test.

3.3 Human-likeness in Terms of Closeness

In this section, metrics 3) **model-human distance**, 4) **closest model-human distance**, and 5) EMD are discussed.

3.3.1 The Model-Human Distance

For **model-human distance**, we perform the two-tailed t-test between every pair of models and the results are in Table 3.4. Also, we perform a t-test between every model and the human participants, but this comparison should be interpreted carefully. First, the distance array we compute on the human side is *not* the average inter-human distance; instead, a human is considered as a model and the average difference to every human participant in the sample, *including him- or herself*, is

computed, before the result is averaged across the humans. This way, the t-test between the model and the humans tells, put intuitively, whether the model prediction has the tendency to stay inside the cloud of human predictions. This result is presented in Table 3.5. Reading both tables, we see:

Conclusion 6. For MPT, Dolly, Falcon, and GPT4, when inter-human divergence is already high, their predictions tend to be even more outrageous, intuitively falling outside the cloud of human predictions. They are not any better than the baseline on this front.
Conclusion 7. As long as there is minimal inter-human disagreement (from the second split on), SpeechT5's predictions tend to fall inside the cloud of human predictions and, in fact, be as close to golden predictions as the humans' (see Tables 3.1 and 3.2).

Split	**Full Ranking**
avg. inter-human dist. $= 0$	SpeechT5 $>$ GPT5 = Gemini2 = Gemini3 = kbG2P = myG2P $>$ GPT4 = Llama3 = Mistral = Qwen2 $>$ Dolly = Falcon = MPT $>$ baseline
$\in (0, 2]$	GPT4 = GPT5 = Gemini2 = Gemini3 = Qwen2 = SpeechT5 = kbG2P = myG2P $>$ Llama3 = MPT = Mistral $>$ Dolly = Falcon = baseline
$\in (2, 5]$	GPT5 = Gemini2 = Gemini3 = Llama3 = Mistral = Qwen2 = SpeechT5 = kbG2P = myG2P $>$ Dolly = Falcon = GPT4 = MPT $>$ baseline
> 5	Gemini2 = Gemini3 = Llama3 = Mistral = SpeechT5 = kbG2P = myG2P $>$ GPT4 = GPT5 = MPT = Qwen2 = baseline $>$ Dolly = Falcon
overall	Gemini2 = Gemini3 = SpeechT5 $>$ GPT4 = GPT5 = Llama3 = Qwen2 = kbG2P = myG2P $>$ Mistral $>$ Dolly = Falcon = MPT $>$ baseline

Table 3.4: Ranking the models by the average model-human distance. Pair-wise comparison is based on t-test and Kahn's algorithm is applied to give the full ranking. The smaller the distance, the higher the ranking is. Models in the same tier (connected by =) are not significantly different on the t-test.

	avg. inter-human dist. $= 0$		$\in (0, 2]$		$\in (2, 5]$		> 5		**overall**	
Model	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p	Mean Diff.	p
Llama3	1.336	**0.000**	1.130	**0.008**	0.315	0.414	0.707	0.279	1.103	**0.000**
Mistral	1.448	**0.000**	1.403	**0.003**	0.699	0.158	0.759	0.078	1.275	**0.000**
Qwen2	1.530	**0.000**	0.931	**0.013**	0.218	0.490	0.932	**0.037**	1.191	**0.000**
Gemini2	0.745	**0.000**	0.692	**0.019**	***-0.014***	0.969	0.172	0.386	0.580	**0.000**
GPT4	1.228	**0.000**	0.838	**0.011**	1.003	**0.033**	0.388	**0.025**	1.051	**0.000**
MPT	2.253	**0.000**	1.979	**0.000**	1.904	**0.004**	1.396	**0.017**	2.078	**0.000**
Dolly	2.851	**0.000**	2.575	**0.000**	0.926	**0.024**	1.043	**0.005**	2.375	**0.000**
Falcon	2.849	**0.000**	2.515	**0.000**	1.591	**0.005**	1.300	**0.003**	2.477	**0.000**
Gemini3	0.823	**0.000**	1.106	**0.021**	0.141	0.671	***-0.086***	0.544	0.698	**0.000**
GPT5	0.993	**0.000**	1.259	**0.005**	0.267	0.451	0.619	0.086	0.908	**0.000**
SpeechT5	0.343	**0.002**	0.442	0.074	0.302	0.346	0.711	0.154	0.389	**0.000**
kbG2P	1.206	**0.000**	0.670	0.068	***-0.186***	0.502	0.246	0.375	0.834	**0.000**
myG2P	1.008	**0.000**	0.686	0.056	0.659	0.130	0.520	0.120	0.859	**0.000**
baseline	3.513	**0.000**	3.585	**0.000**	2.769	**0.001**	0.858	**0.021**	3.182	**0.000**

Table 3.5: Two-tailed t-test that compares the average model-human distance and the average human-human distance (where the zero distance between a human and themselves is included). The model mean or human mean is ≥ 0, so the difference is in theory unbounded. Intuitively, the lower it is, the closer the model predictions are to the "geometric median" of the human predictions.

3.3.2 The Closest Model-Human Distance

We present two results in this part: the average modal-human distance for each model (Table 3.6) and the ranking of all models (based on the t-test) (Table 3.7).

To provide some context for interpreting Table 3.6, the distance between [t] and [ʧ] - two consonant phones with the same place of articulation but different manners - is 1.5, and the distance between [i] and [ə] - one high-front vowel sound and one centre vowel sound - is 1.25. A minimal pair like [t] and [d], who differ by just one feature, has distance 0.25. Long and short vowel sounds [iː] and [i] has distance 0.5. We consider the final two cases as "phonetically close". By this standard, we can tell:

Conclusion 8. On average, all LLMs fail to give a prediction that is phonetically close to *any* human prediction, an achievement that only SpeechT5 more or less managed.

The comparisons made in table 3.7 further indicates:

Conclusion 9. The two G2P models are hardly surpassed by, if not better than, LLMs. Such models as Llama3, Mistral, Qwen2, GPT4, MPT, Dolly, and Falcon all fall behind the G2P models on the front of giving a prediction that is conceivable to humans.

Model	**avg. inter-human dist.** $= 0$	$\in (0, 2]$	$\in (2, 5]$	> 5	**overall**
Llama3	1.336	1.228	0.405	1.173	1.176
Mistral	1.448	1.498	0.629	1.556	1.356
Qwen2	1.530	0.968	0.838	1.505	1.334
Gemini2	0.745	0.750	0.451	0.431	0.677
GPT4	1.228	0.851	0.930	0.827	1.084
MPT	2.253	1.974	1.228	1.676	2.011
Dolly	2.851	2.635	1.377	1.439	2.484
Falcon	2.849	2.581	1.523	1.733	2.519
Gemini3	0.823	1.179	0.453	0.365	0.794
GPT5	0.993	1.291	0.598	1.033	0.997
SpeechT5	0.343	0.495	0.214	1.122	0.424
kbG2P	1.206	0.764	0.286	0.553	0.943
myG2P	1.008	0.819	0.652	1.151	0.939
baseline	3.513	3.652	2.529	1.437	3.214

Table 3.6: The average distance between the model prediction and the closest human prediction.

Split	**Full Ranking**
avg. inter-human dist. $= 0$	SpeechT5 > GPT5 = Gemini2 = Gemini3 = kbG2P = myG2P > GPT4 = Llama3 = Mistral = Qwen2 > Dolly = Falcon = MPT > baseline
$\in (0, 2]$	GPT4 = GPT5 = Gemini2 = Gemini3 = Qwen2 = SpeechT5 = kbG2P = myG2P > Falcon = Llama3 = MPT = Mistral > Dolly = baseline
$\in (2, 5]$	GPT5 = Gemini2 = Gemini3 = Llama3 = Mistral = Qwen2 = SpeechT5 = kbG2P = myG2P > Dolly = Falcon = GPT4 = MPT > baseline
> 5	GPT5 = Gemini2 = Gemini3 = Llama3 = SpeechT5 = kbG2P = myG2P > GPT4 = Mistral = Qwen2 = baseline > Dolly = Falcon = MPT
overall	Gemini2 = SpeechT5 > GPT5 = Gemini3 = kbG2P = myG2P > GPT4 = Llama3 = Mistral = Qwen2 > MPT > Dolly = Falcon > baseline

Table 3.7: Ranking the models' human-likeness by the average distance to the closest human prediction. Pair-wise comparison is based on t-test and Kahn's algorithm is applied to give the full ranking. The smaller the distance, the higher the ranking is. Models in the same tier (connected by =) are not significantly different on the t-test.

3.3.3 The Earth Mover's Distance (EMD)

First, we present two results in this part: the average EMD between the models predictions and the humans predictions (Table 3.8) and the ranking of all models (based on the t-test) (Table 3.9). Note that, since EMD assigns the total probability of 1 to each prediction of the model/humans - meaning the EMD between ([i]) and ([ə]) is equal to the EMD between ([i], [i]) and ([ə], [ə]) -, the sample distances mentioned earlier can still serve as reference. An EMD of 1.5 is considered not phonetically close. This time, when inter-human divergence is high (the final two splits), even SpeechT5's predictions are not close to humans', no to mention other models.

Conclusion 10. LLMs' predictions, considered as a distribution, are not phonetically close to human predictions.

Model	**avg. inter-human dist.** $= 0$	$\in (0, 2]$	$\in (2, 5]$	> 5	**overall**
Llama3	1.336	1.597	2.881	4.418	1.875
Mistral	1.448	1.874	3.162	3.962	1.987
Qwen2	1.530	1.404	2.414	4.023	1.856
Gemini2	0.745	1.148	2.557	3.522	1.317
GPT4	1.228	1.283	3.380	3.262	1.716
MPT	2.253	2.474	4.489	4.834	2.831
Dolly	2.851	3.004	3.141	4.496	3.069
Falcon	2.849	2.964	3.867	4.596	3.168
Gemini3	0.823	1.607	2.742	3.616	1.479
GPT5	0.993	1.721	2.701	4.317	1.660
SpeechT5	0.343	0.875	2.831	4.196	1.128
kbG2P	1.206	1.210	2.601	4.314	1.681
myG2P	1.008	1.226	3.446	4.588	1.706
baseline	3.513	4.125	5.556	4.926	4.029

Table 3.8: The average Earth Mover's Distance (or Wasserstein distance) between all model predictions and all human predictions.

Next, we recall that Speech T5 did manage to be close to humans in terms of the **closest model-human distance** on the third split ($\in (2, 5]$). In fact, if we compare Tables 3.2, 3.6, 3.8, we find that in Table 3.6, almost all models have the lowest **closest model-human distance** on the third split; meanwhile, their performance in terms of the **distance to golden** and their distributional human-likeness in terms of the **EMD** are not the best among the four splits. Put together, the possible implication is:

Hypothesis 11. On - and only on - pseudo-words that are difficult enough but not impossibly difficult, model predictions tend to get closer to some *unpopular* human predictions.

However, this is not yet statistically supported. To see if there is a sweet spot, which seems to be

where the inter-human distance is $\in (2, 5]$, for models in terms of the **closest model-human distance**, we perform a quadratic regression. We also perform the Pearson correlation test to examine not only the linear correlation between the inter-human distance and and the *closest model-human distance*, but also that between the inter-human distance and the **EMD**. The results are in Figures 3.3 & 3.4, 3.5 & 3.6, and 3.7 & 3.8 respectively. In each pair of figures connected by "&", one excludes the first split, where each of the 123 pseudo-words has inter-human distance = 0, and the other includes all 4 splits.

Split	**Full Ranking**
avg. inter-human dist. $= 0$	SpeechT5 > GPT5 = Gemini2 = Gemini3 = kbG2P = myG2P > GPT4 = Llama3 = Mistral = Qwen2 > Dolly = Falcon = MPT > baseline
$\in (0, 2]$	GPT4 = GPT5 = Gemini2 = Gemini3 = Qwen2 = SpeechT5 = kbG2P = myG2P > Llama3 = MPT = Mistral > Dolly = Falcon = baseline
$\in (2, 5]$	Dolly = GPT4 = GPT5 = Gemini2 = Gemini3 = Llama3 = Mistral = Qwen2 = SpeechT5 = kbG2P > Falcon = MPT = myG2P > baseline
> 5	GPT4 = Gemini2 = Gemini3 = Llama3 = Mistral = Qwen2 = SpeechT5 > Dolly = Falcon = GPT5 = MPT = baseline = kbG2P = myG2P
overall	Gemini2 = SpeechT5 > GPT4 = GPT5 = Gemini3 = Qwen2 = kbG2P = myG2P > Llama3 = Mistral > Dolly = Falcon = MPT > baseline

Table 3.9: Ranking the models by the Earth Mover's Distance (or Wasserstein distance) between all model predictions and all human predictions. Pair-wise comparison is based on t-test and Kahn's algorithm is applied to give the full ranking. The smaller the distance, the higher the ranking is. Models in the same tier (connected by =) are not significantly different on the t-test.

While Figures 3.5 & 3.6 make it clear that there is no correlation between a model's ability to given some conceivable prediction and the difficulty of the pseudo-word - there is, in fact, a tendency of negative correlation in "weaker" models like MPT, Dolly, and Falcon[4]- there is also no quadratic

[4]This is understandable. Since they are "weaker" performance-wise, they tend to stay no where near the cluster(s) of human predictions when humans show strong agreement; on harder pseudo-words, however, human predictions are scattered about the metric space, and these models' no-where-near-reasonable predictions have a better chance of finding a close by human guess.

relation, as shown in Figures 3.3 & 3.4 (the quadratic p-value is the primary statistic, telling how likely modeling the data with a parabola is delusional). That is, although Table 3.6 says that the third split (the second hardest one) sees nearly all models being the closest to *any* human prediction at the group average level, this tendency is not seen at the point level. However, since the trend in Table 3.6 is very clear, this phenomenon requires further investigation.

On the other hand, Figures 3.7 & 3.8 make us very confident to say that the models (LLMS, SpeechT5, and G2P models) tend to give a distribution of predictions that is closer to the distribution of human predictions when the inter-human distance goes low. The only exceptions are GPT5, Dolly, and Falcon when the first split is excluded. Though not surprising, this sends an important message: while the distribution of model predictions is not very human-like (**Conclusion 10**), still we have:

Conclusion 12. In terms of the distribution of predictions, LLMs are more human-like when the pseudo-word is easier.

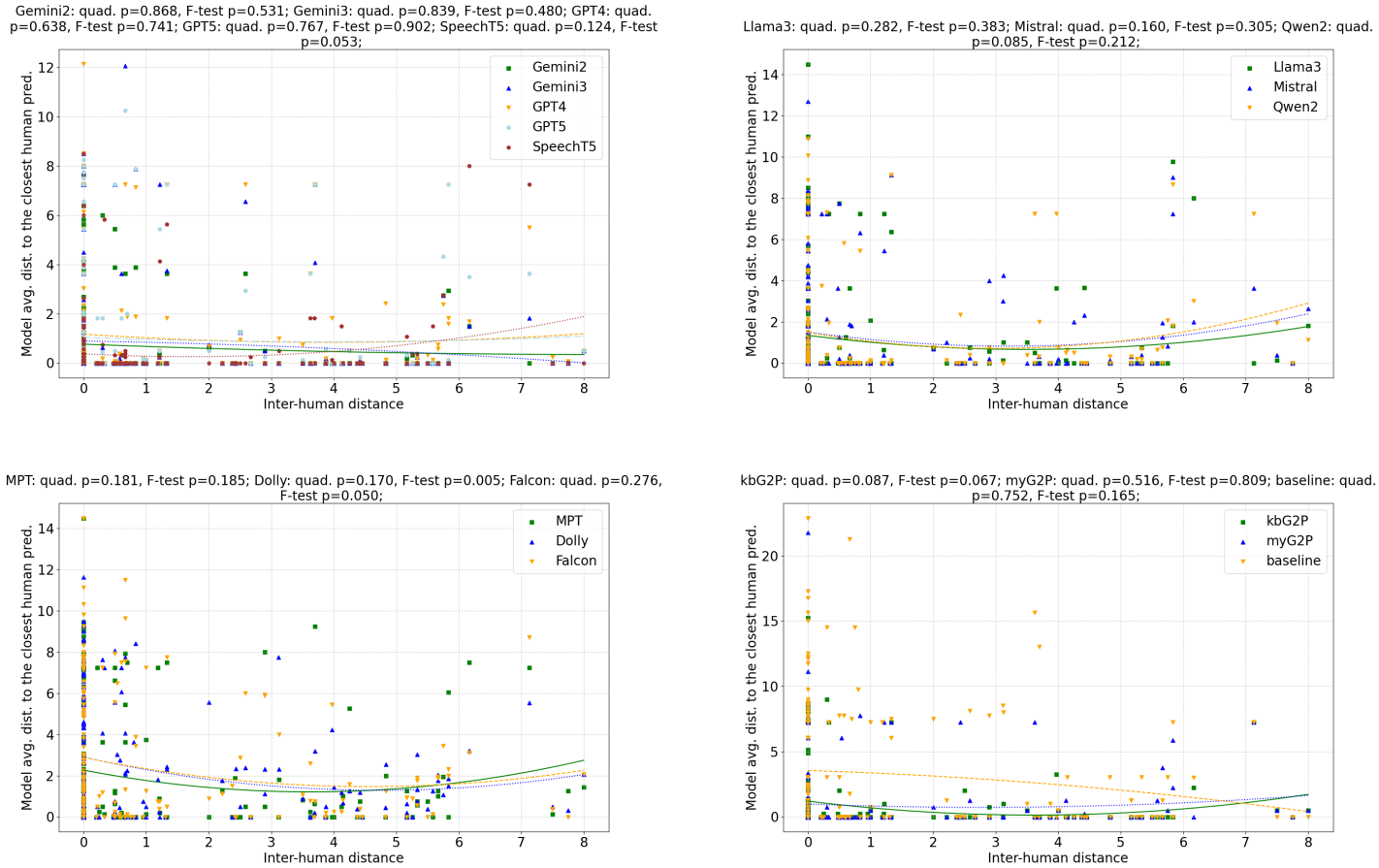


Figure 3.2: Quadratic Regression. The X-axis is the inter-human distance of a pseudo-word. The Y-axis is the average model distance to the closest golden prediction, indicating the model's ability to give some conceivable prediction. This plot includes all 4 splits.

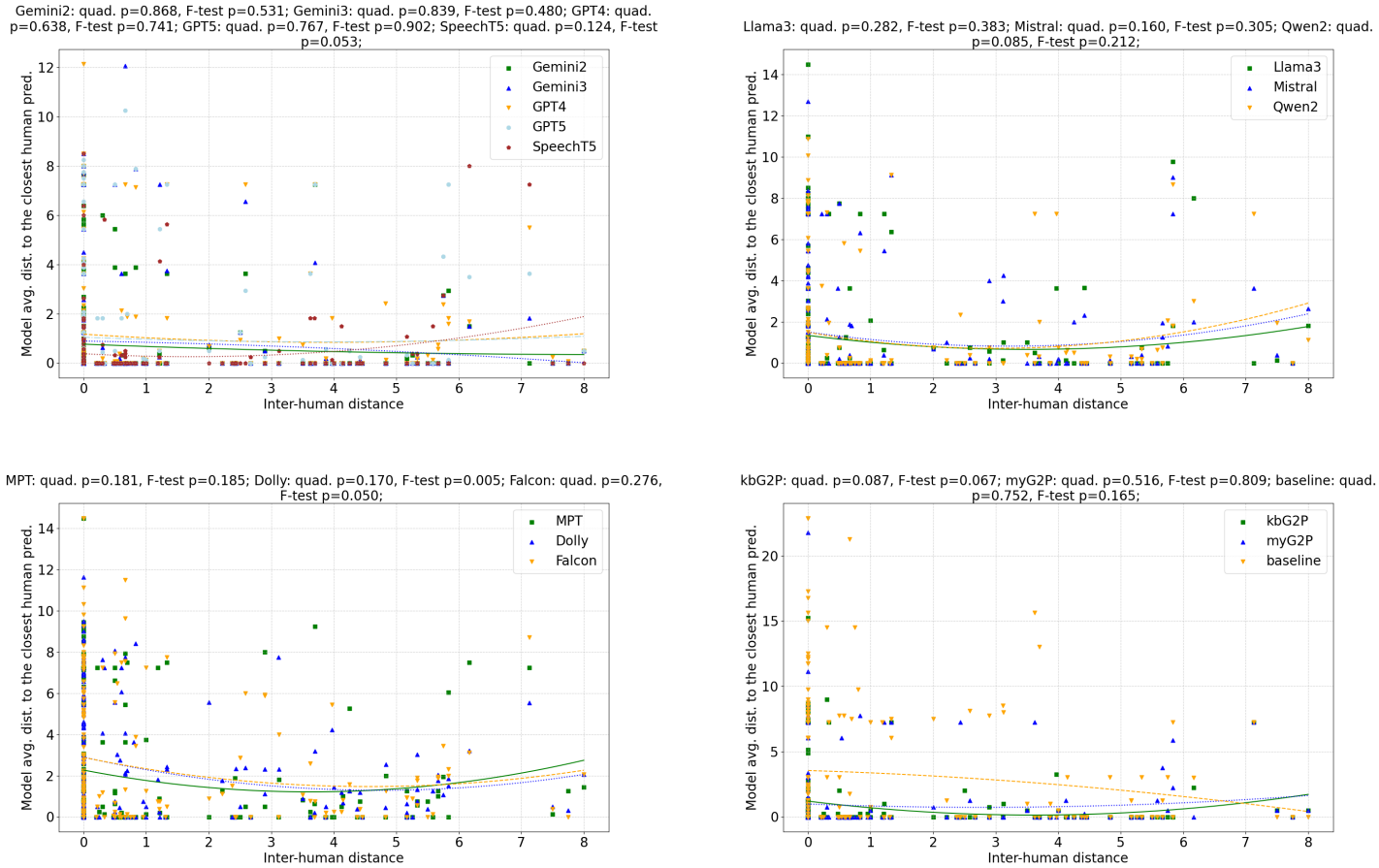


Figure 3.3: Quadratic Regression. The X-axis is the inter-human distance of a pseudo-word. The Y-axis is the average model distance to the closest golden prediction, indicating the model's ability to give some conceivable prediction. This plot includes all 4 splits.

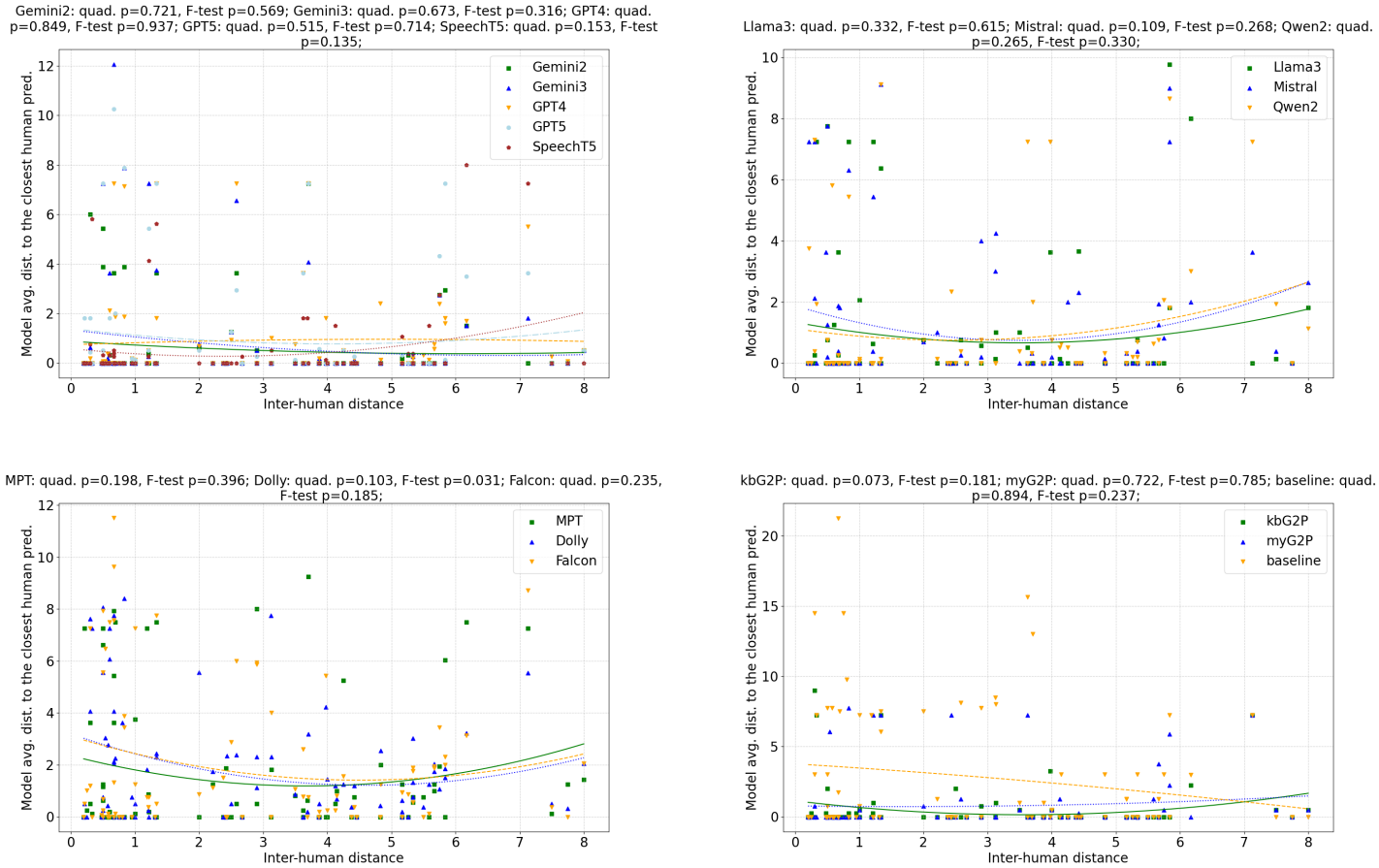


Figure 3.4: Quadratic regression. The X-axis is the inter-human distance of a pseudo-word. The Y-axis is the average model distance to the closest golden prediction, indicating the model's ability to give some conceivable prediction. This plot excludes the first split (inter-human distance = 0).

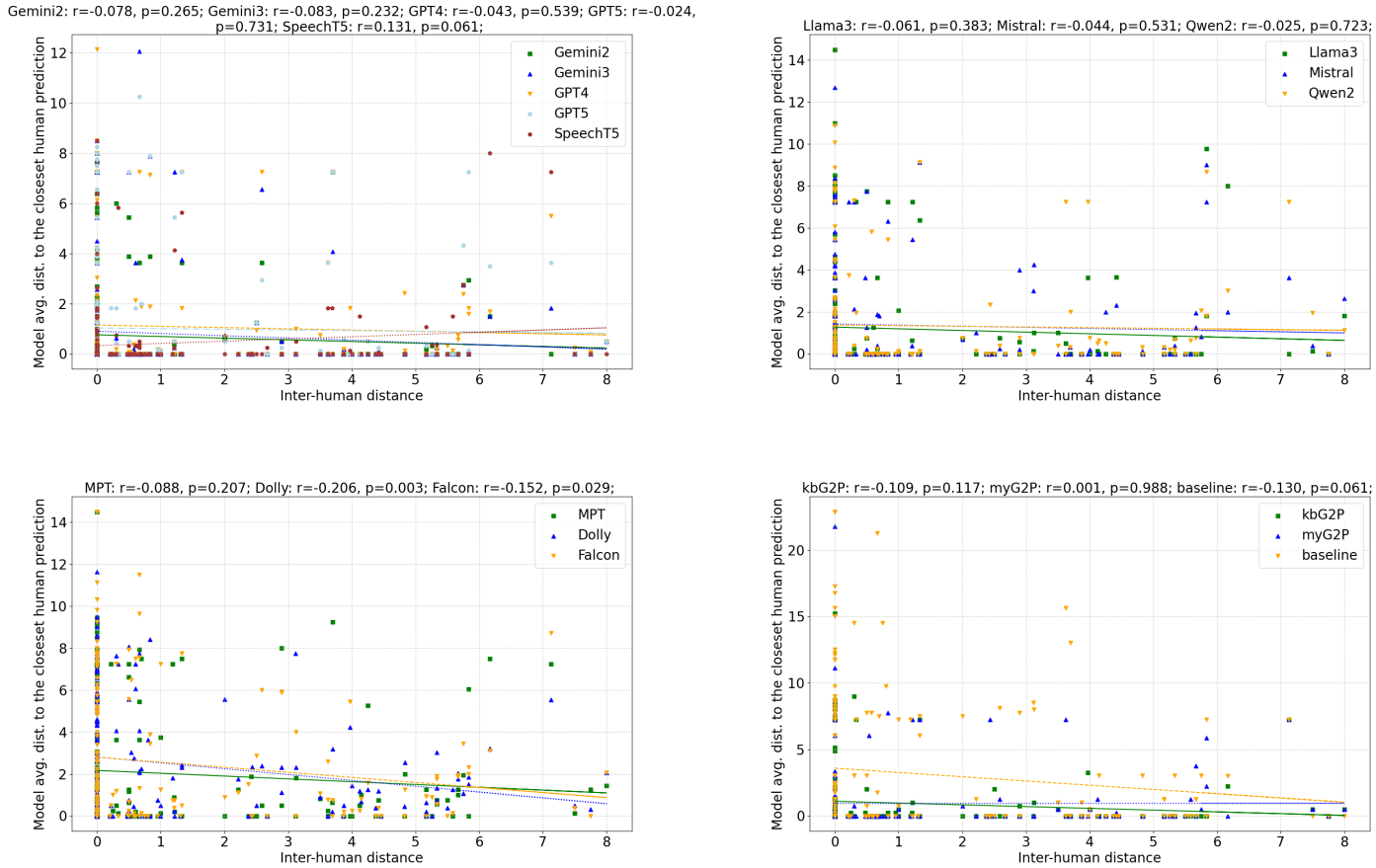


Figure 3.5: Linear Correlation. The X-axis is the inter-human distance of a pseudo-word. The Y-axis is the average model distance to the closest golden prediction, indicating the model's ability to give some conceivable prediction. This plot includes all 4 splits.

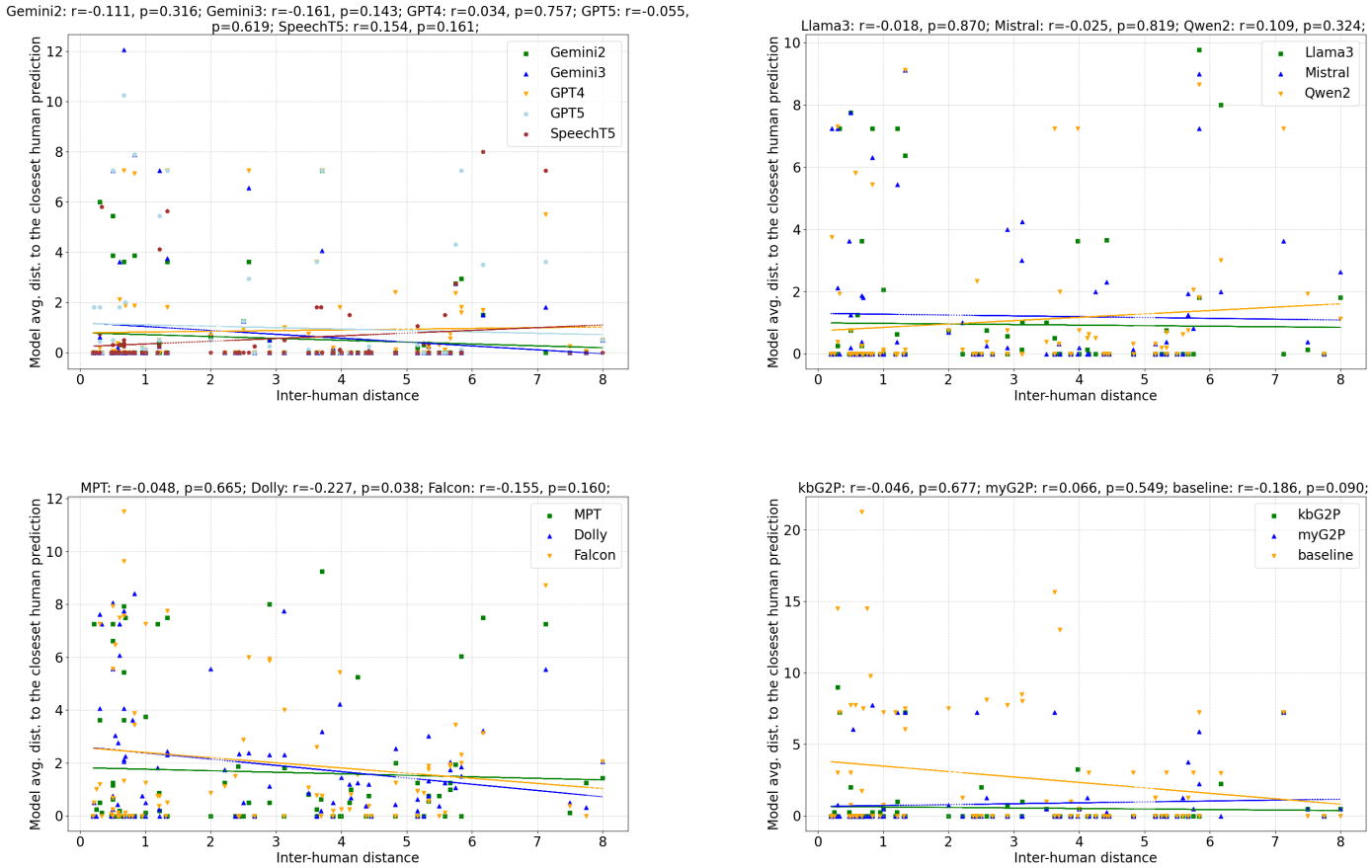


Figure 3.6: Linear Correlation. The X-axis is the inter-human distance of a pseudo-word. The Y-axis is the average model distance to the closest golden prediction, indicating the model's ability to give some conceivable prediction. This plot excludes the first split (inter-human distance = 0).

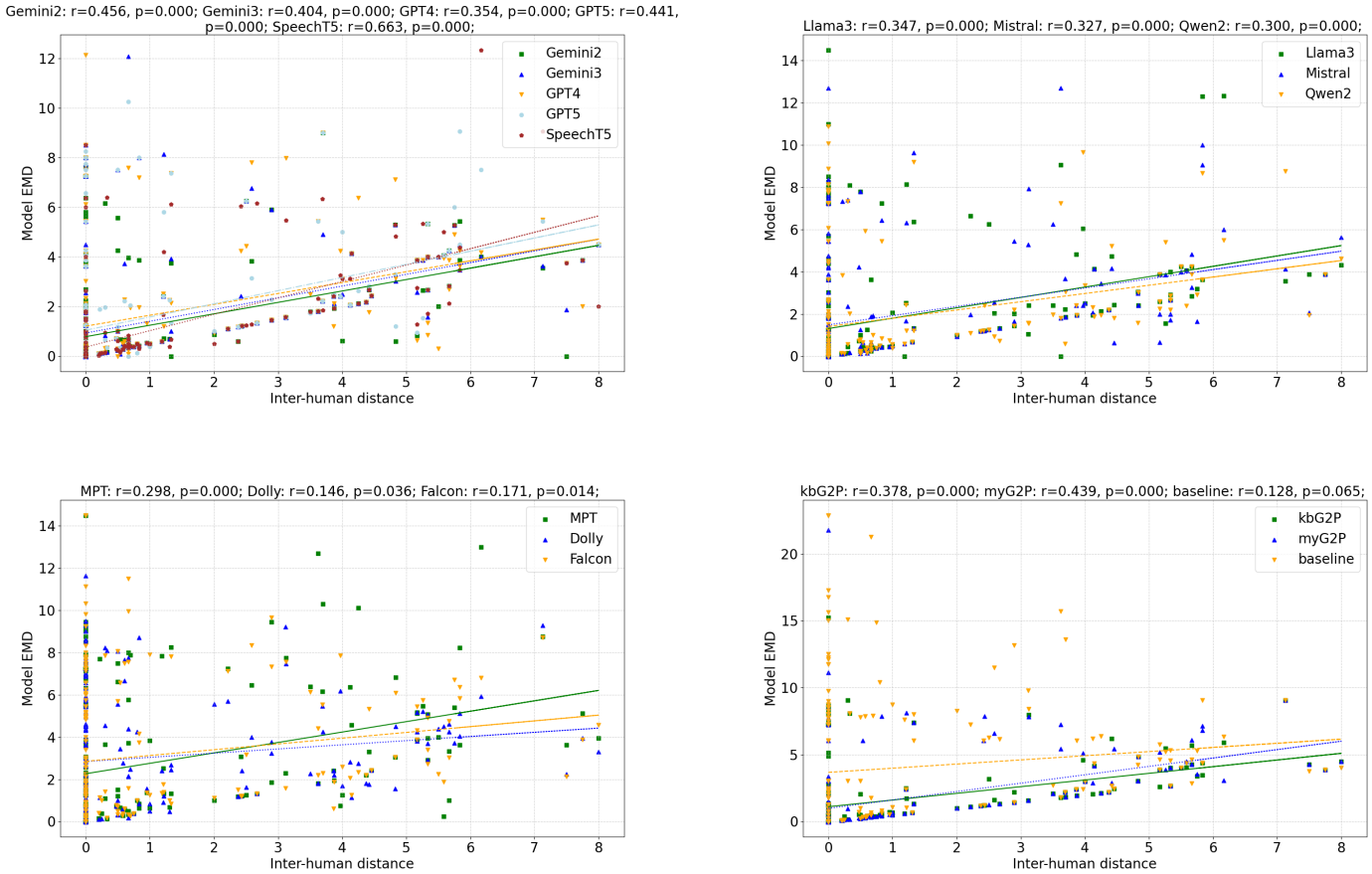


Figure 3.7: The X-axis is the inter-human distance. The Y-axis is the EMD between the model predictions and the human predictions. The plot includes all four splits.

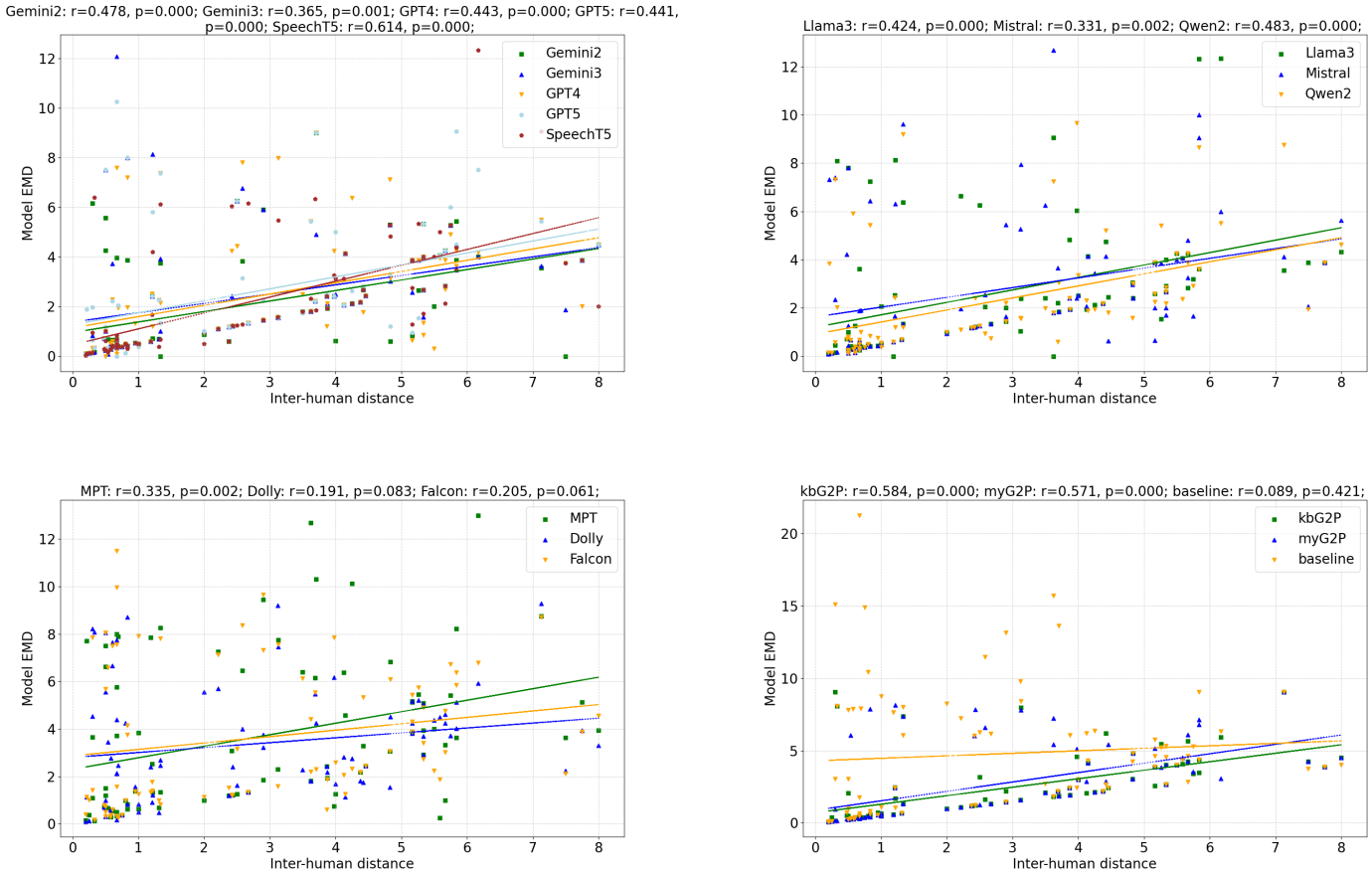


Figure 3.8: The X-axis is the inter-human distance. The Y-axis is the EMD between the model predictions and the human predictions. This plot excludes the first split (inter-human distance = 0).

3.4 Human-likeness in Terms of Feeling the Difficulty

At last, we examine whether LLMs and humans feel the difficulty of pseudo-words the same way. This can be broken further down to two questions:

1) If humans are bewildered (high **inter-human distance**) upon seeing a pseudo-word, are models also bewildered (high inter-run distance) at this pseudo-word?

2) If humans underperform (high **distance to golden**) on a pseudo-word, do models also underperform (high **distance to golden**) here?

The first two questions are addressed by Figures 3.9 & 3.10 and 3.11 & 3.12 respectively, where the Pearson correlation test is carried out. In each pair of figures connected by "&", one excludes the first split, where each of the 123 pseudo-words has inter-human distance = 0, and the other includes all 4 splits.

Figures 3.9 & 3.10 show that, when humans are bewildered, so are Gemini 3, GPT4, Mistral, and Qwen2, but not other LLMs - including Gemini2, GPT5, and Llama3. The fact actually rejects the high-level idea that an LLM that makes more human-like predictions is also more human-like in terms of bewilderment. Actually, this suggests that LLMs' human-like performance does not reflect human-like reasoning process, which is more or less in line with researches showing models' (not only LLMs) reasoning might be decorative or at least non human-like (see [4], [58], for example). SpeechT5 also exhibits strong correlation. Note that the baseline and G2P models are irrelevant here as they are deterministic and were run once only.

Figures 3.11 show that, with all splits considered, all models (LLMs, SpeechT5, and G2P models), are correlated with human performance in terms of the distance to the closest golden prediction. Considering that little correlation is spotted for the baseline, we can tell at a high level that there is really certain human-likeness in these neural network models. In Figure 3.12, when the first split is excluded, Gemini3, GPT4, Llama3 and Mistral no longer exhibit strong correlation ($p \geq 0.05$ now), implying their overall performance is much benefited from the easiness of the first split, but not, say, the second.

Put together, the results suggest at a high-level:

Conclusion 13. LLMs do not necessarily feel the difficulty of a pseudo-word the human way in terms of bewilderment, but are highly human-like in terms of performance (assuming there are golden predictions).

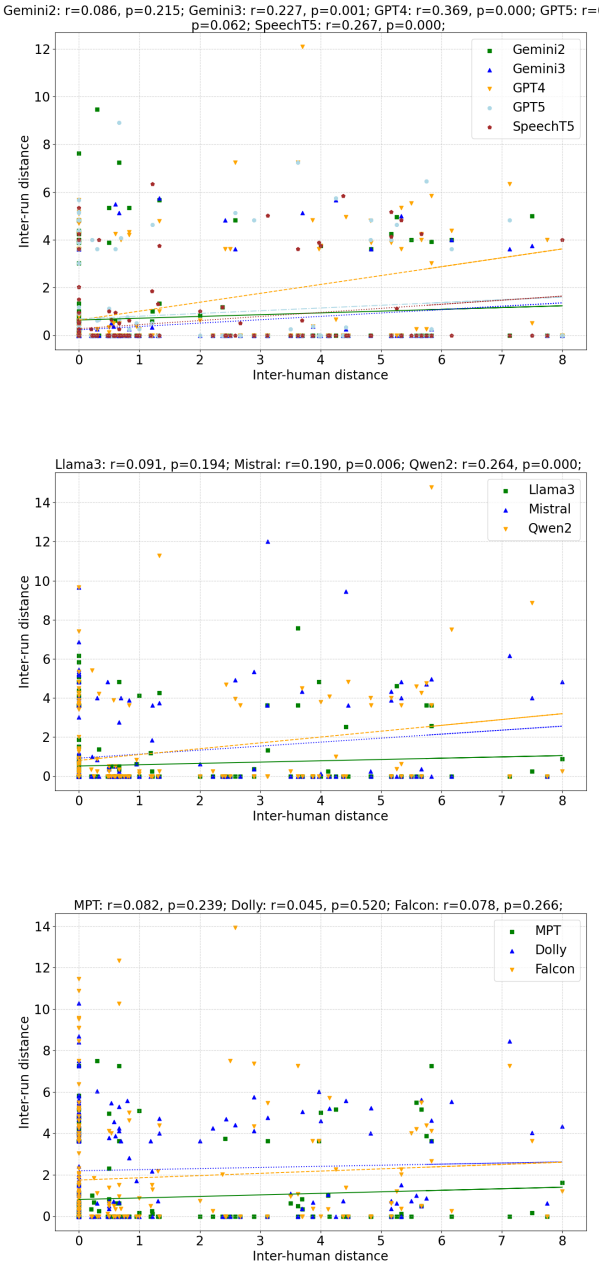


Figure 3.9: The X-axis is the inter-human distance of a pseudo-word, indicating how much humans are "bewildered" upon seeing a pseudo-word. The Y-axis is the inter-run distance, indicating how much the model is "bewildered". This plot includes all 4 splits.

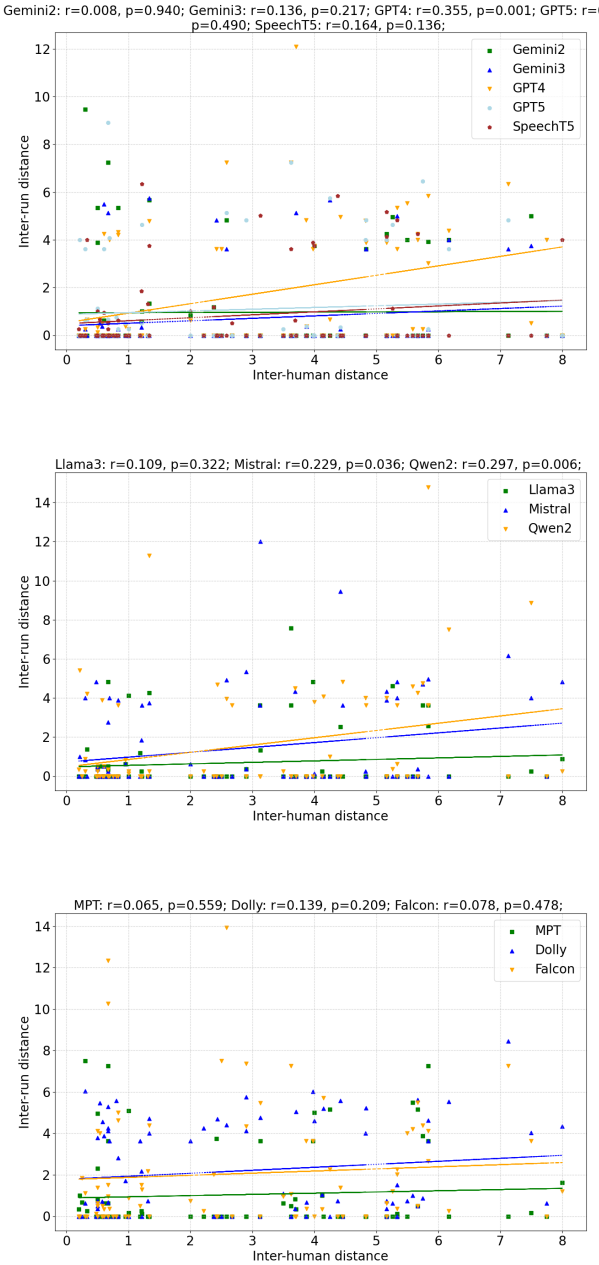


Figure 3.10: The X-axis is the inter-human distance of a pseudo-word, indicating how much humans are "bewildered" upon seeing a pseudo-word. The Y-axis is the inter-run distance, indicating how much the model is "bewildered". This plot excludes the first split (inter-human distance = 0).

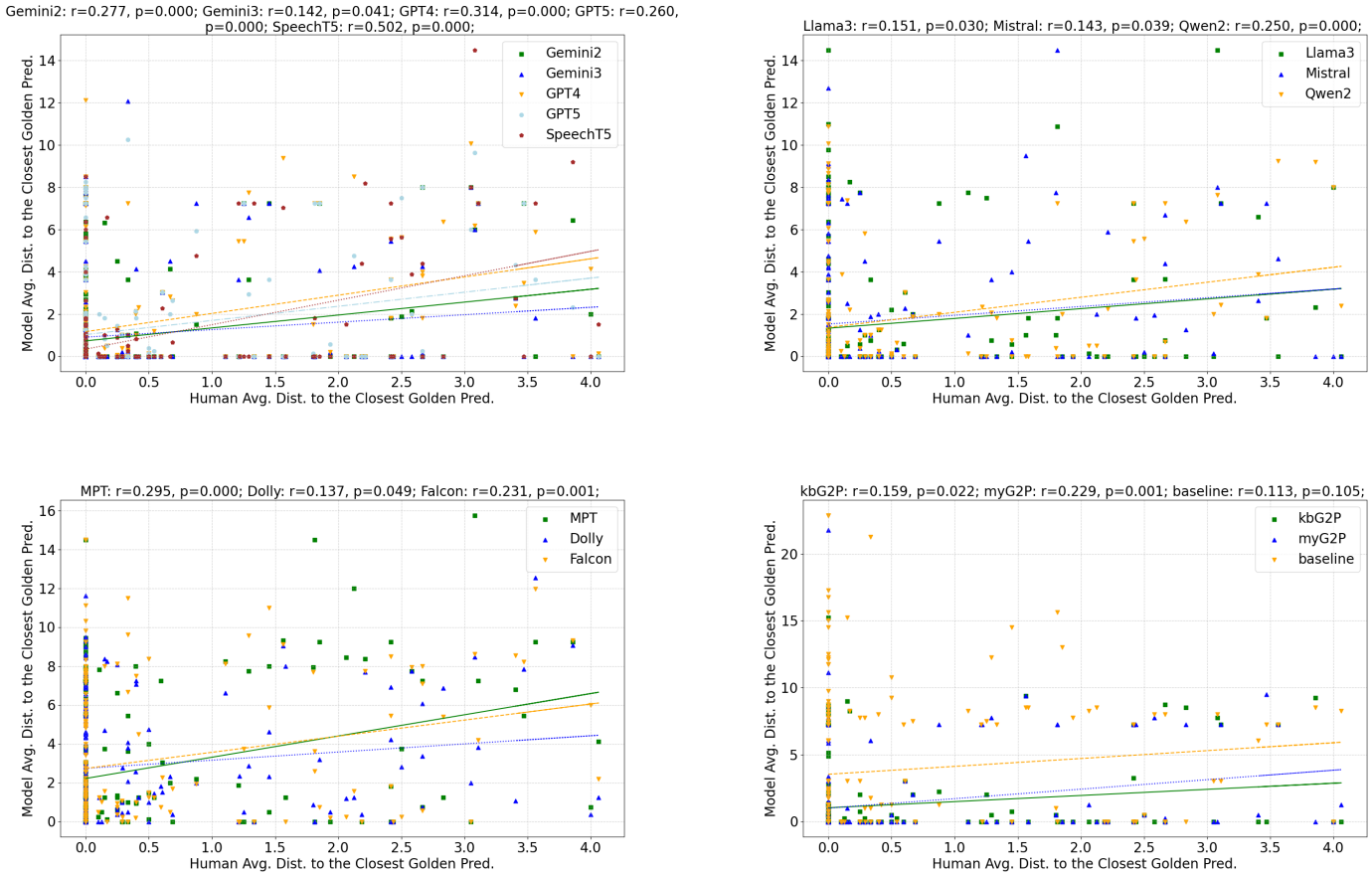


Figure 3.11: The X-axis is the humans' average distance to the closest golden prediction. The Y-axis is the model's average distance to the closest golden prediction. The plot includes all four splits.

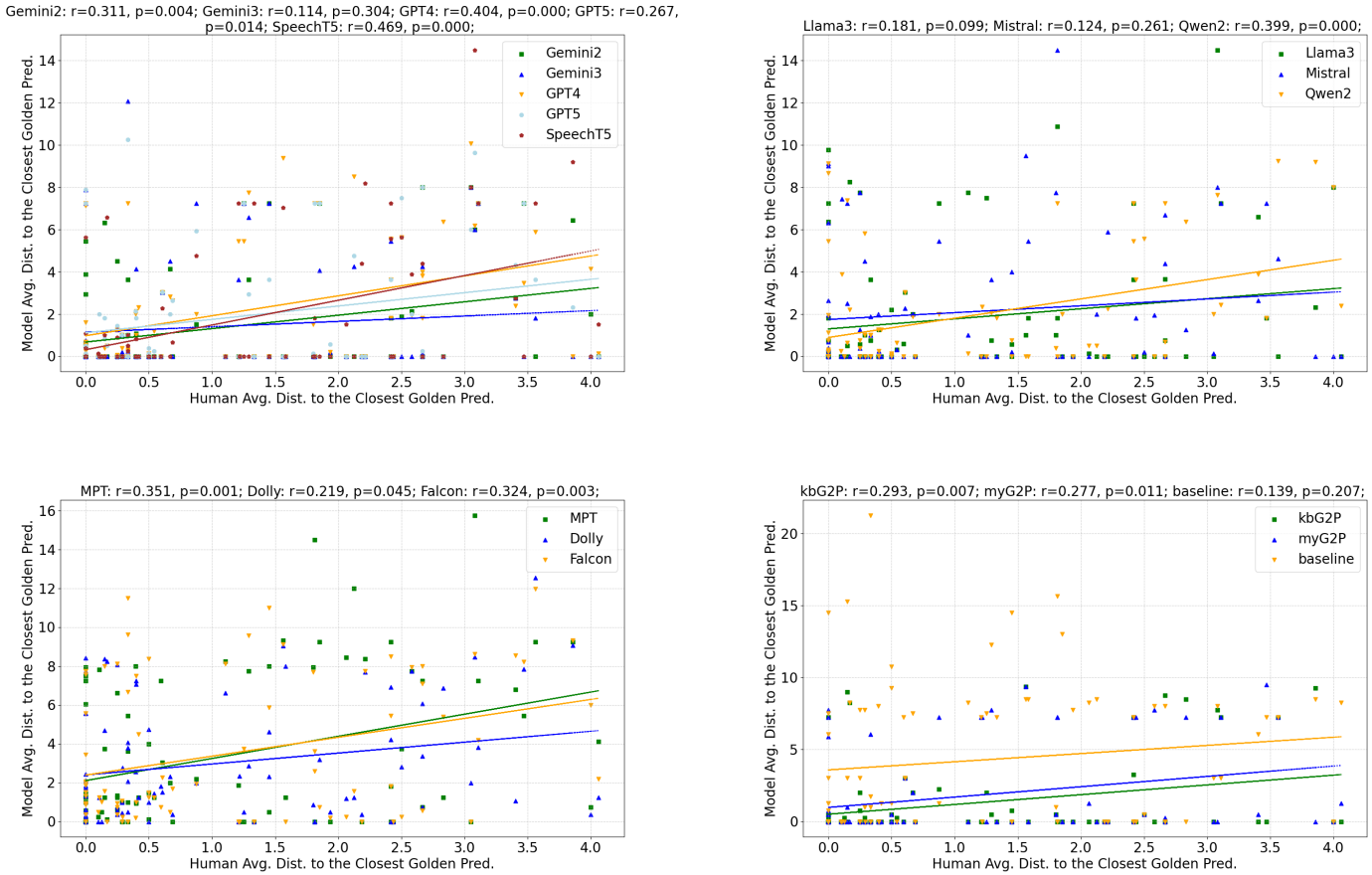


Figure 3.12: The X-axis is the humans' average distance to the closest golden prediction. The Y-axis is the model's average distance to the closest golden prediction. This plot excludes the first split (inter-human distance = 0).

3.5 Discussion

3.5.1 Human-likeness

Recall that our ultimate question is whether LLMs understand English pronunciation and G2P rules the way humans do. With the results we attained so far, we can say the answer to a great extent is "yes". For one thing, LLMs are in general making reasonable attempts, the best ones hitting the golden prediction about 80% of the time on the easiest split (Table 3.1). Moreover, they feel the difficulty of the pseudo-words the similar way humans do given the correlation of performance; such correlation is missing between the baseline and humans (Figures 3.9 & 3.10). Some LLMs' inter-run distances, indicating the bewilderment, are also correlated with the inter-human distance. As for the rest LLMs that fail to exhibit such correlation, the reason might have to do with your methods, which we discuss now.

Our results show that (1) The EMD between model predictions and human predictions is in general ≥ 1.25, which is not considered "phonetically close" and (2) some LLMs' inter-run distances are not correlated with inter-human distances. While they suggest some non-human-likeness of the LLMs, we must admit there is imperfection with our method. In theory, to get the distribution of predictions of a given model, one should, empirically, turn up the temperature and run the model with the same prompting enough times, or, more directly, retrieve the logits from the output layer. Though the latter is the ideal way, in our study it is not easy to put into practice, because we cannot predict which token(s) will contain the sound of the part of interest (POI) and, even for a same pseudo-word, the token order varies depending on the predictions before the POI, making the logits irrelevant to the probabilities we care about. The former empirical way is valid, but unaffordable in terms of the workload. The best we can do now is turn the temperature to zero and use 4 prompt templates, so the variation is actually induced by prompting[5]. As a result, it is unclear if we can assume the predictions made in 4 runs really represent the distribution of model predictions. Computing both the EMD and the inter-run distance are based on this assumption. Therefore, we avoid making strong conclusions based on the two results discussed here.

Actually, LLMs' non-human-likeness is suggested by the very same result mentioned above - on the easiest split, the **hit golden** rate is *only* around 80%, alongside the fact that in general model predictions are not phonetically close to the golden predictions. So the complete story may be: LLMs do understand how English G2P works and can decide in the humans' way whether a pseudo-word is easy or hard for prediction, but their knowledge about English pronunciation and G2P do not completely overlap with the (implicit) knowledge of native human speakers. By "not completely overlap", we imply part of humans' knowledge is beyond LLMs *(lack of knowledge)* and/or part of LLMs' knowledge is beyond humans' knowledge *(wrong knowledge)*.

3.5.2 Comparing Models

If we consider all models involved here as a type of intelligence, we note that apart from MPT, Dolly, and Falcon (the weakest tier), all other LLMs do not show significant difference in terms of

[5]In theory, such randomness is still valid as long as the number of prompt templates is large enough. This is indeed why we use 4 prompt templates, but 4 is obviously too small a count.

both performance (Tables 3.3, 3.7, 3.9) and correlation (Figures 3.9 & 3.10, 3.11 & 3.12). Most remarkably, Gemini2 is not below Gemini3 and GPT4 is not below GPT5 performance-wise, despite the size difference and the generation gap. In terms of EMD and **closest model-human distance**, Gemini2 is even more human-like overall than Gemini3; in terms of correlation, GPT4 is somehow the most human-like LLM (highest r value) throughout. Moreover, the two G2P models are not much worse than LLMs, if not better, in terms of both performance and correlation. On the other hand, SpeechT5 is the most human-like on all aspects, being in the top tier in every ranking, exclusively so on the first split. As a result, it has the highest r value for the correlation of performance.

The overall message seems to be: Given the G2P task, text-to-text models such as MPT, Dolly, and Falcon have room for improvement, but once the level of GPT4, Gemini2, Llama3, etc. is reached, that is it. This level is also reachable for good enough BiLSTM G2P models. Larger model size and training corpus offers little help. If we aim for more human-like predictions, at least on the easiest split like humans do, SpeechT5's results suggests that we probably have no choice but to overhaul the model nature and/or the training modality. Our philosophy of the LLM is that knowledge, be it of languages, logic, or math, is learnt through text reading, as it can be learnt by humans through this modality. In theory, phonetic/phonological and G2P knowledge can be learnt this way, too, provided the sound of the entire English vocabulary is available inside the training corpus[6] - despite the fact that humans, native speakers in particular, do not learn it this way. It turns out that SpeechT5, representing the models whose training involves both sound and text, better resembles humans. It is worth exploring if there are indeed inherent limitations to LLMs - or text-to-text models in general - on this front.

[6]This is exactly how the two G2P models learn, and this is far from the best education an LLM can receive. In fact, better LLMs such as Gemini3 seem to be able to cite some G2P rules when asked for justification - and without actually applying the rules when making the prediction in the same response.

Chapter 4
Analysing Tendencies

One key message from the last chapter is that LLMs are making human-like attempts but their performance falls short of that of humans - if we assume golden predictions reflect some truth about humans' implicit knowledge of English pronunciation and G2P. This chapter looks at LLMs' mistakes closer up and basically addresses this question: Do LLMs have any specific tendencies when predicting the sound such that their predictions are deviate from the golden prediction among human participants? Given the fact that MPT, Dolly, and Falcon prove to be weaker on the current task, their deviation from humans being more frequent and possibly coming in different patterns, we exclude them in this part. To better study the tendency of top tier LLMs, the examination is **limited to Llama3, Mistral, Qwen2, Gemini2, Gemini3, GPT4, and GPT5** - recall from the previous chapter that the 7 models are not significantly different on most metrics.

The tendencies, unsurprisingly, turn out to be manifold, and we discuss them in each subsection. The first three have to do with predicting consonants (subsections 4.1.1-3.) and the rest five predicting vowels (subsections 4.2.1-6.).

One methodological concern is whether the spotted tendency was more or less induced by the specific set of pseudo-words we created. To address this concern, we sample a set of words from the real English vocabulary and ask Llama3 and Mistral to predict the pronunciation. In fact, if an existing English word is given to Genimi2 and GPT4, not to mention Gemini3 and GPT5, the model is always able to provide the correct answer, likely memorised from its training data. This, however, does not happen with Llama3, Mistral, and Qwen2. While we believe such pronunciations are also included in the pre-training data (like Wikitionary and crawl-able online dictionaries) of the last three, they fail to leverage the information due to model reasons, data reasons, or training reasons. Nevertheless, this does not mean the three have little opportunity to learn about English G2P[1], which is suggested by their performance on pseudo-words. They are just not as good at memorising facts available in the pre-training data as the other four, which makes real English words another group of "pseudo-words" for them. Therefore, we can run validation experiments on the three using real words. Due to limited time, we only work with Llama3 and Mistral, excluding Qwen2 in this study[2]. In most experiments, we also include our baseline algorithm that is based on string co-occurrence in an attempt to show that some mistakes are avoidable even for a frequency-based algorithm. If a tendency seen on pseudo-words also show in the validation experiment, it strongly suggests that LLMs - or at least Llama3 and Mistral - do have the tendency. The prompt looks this way:

[1] Otherwise it can be pointless examining their "knowledge" of English pronunciation and G2P to begin with.

[2] Using real words is also a means to save time and work. Admittedly, the best practice is to create another batch of pseudo-words for validation purpose.

Give the pronunciation of the word [WORD] in IPA in North American English.

4.1 Tendencies with Consonants

First, we list the mistakes (i.e. predictions different from the golden prediction) made by all 7 LLMs in question, sorted by the occurrences (Tables 4.1 and 4.2). A count of 7 roughly means each model has made this mistake once - out of 4 runs - on average, and is considered noteworthy. The last column in the tables provides the pseudo-words where the mistake happens along with the ratio "count of the golden prediction : count of the model prediction" among human participants. For example, "smissor(4:0)" in the first row means 4 human participants hit the golden prediction [z] while 0 human participants give the model prediction, which is [s]. Looking at these ratios in the two tables, we are convinced that LLMs' mistakes are highly non-human-like.

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
1	z	s	33	smi**ss**or(4:0), oo**s**y(3:0)
2	ɹ	wɹ	18	**wr**eta(4:0), **wr**ithlob(3:0)
3	t^h	bt^h	13	nou**bt**(2:0)
4	n	gn	13	di**gn**(5:0), calei**gn**(3:0)
5	n	k^hn	11	**kn**othy(6:0), **kn**eck(4:0))
6	ð	θ	10	mou**th**erdam(4:0)
7	ʒ	z	9	dei**z**ure(5:0)
8	m	mn	8	ma**mn**(7:0)
9	ʧ	k^h	8	**ch**eme(4:1)
10	m	mən	7	ma**mn**(7:1)
11	ʃ	s	7	po**c**ean(2:1)
12	f	p^h	7	liccou**gh**(3:0), te**ph**erd(4:2)

Table 4.1: LLMs' mistakes regarding consonants. This table includes mistakes with count ≥ 7. The last column shows the pseudo-word(s) where the mistake happens with the part of interest in bold, and the numbers in the parentheses are "count of the golden prediction" : "count of the mistake" among human participants.

4.1.1 Rigid Mappings to Fricatives

Rows 1, 6, 7 and 11 in Table 4.1 suggest that LLMs have a strong tendency to stick to rigid mappings such as: "*ss* or *s* must map to [s]", "*th* must map to [θ]", "*c* must map to [s]". While "*ss* or *s*→[s]", "*th*→[θ]", and "*c*→[s]" are more common than "*ss* or *s* →[z]", "*th*→[ð]", and "*c*→[ʃ]" respectively, human participants are able to learn from the English vocabulary that the uncommon can happen and have no difficulty deciding when that should happen. But row 13 in Table 4.2 seems to say LLMs can also follow the uncommon mapping "*th*→[ð]" at times, though in a non-human-like way once again. Note that all these examples have to do with fricatives. To see if LLMs do have the tendency to follow the most common mappings to fricatives, we run the validation experiment on

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
13	θ	ð	6	**th**easond(7:0), fy**th**(3:0)
14	kw	k	6	vas**q**uido(4:2)
15	nj	p^hn	6	**pn**eudob(3:0)
16	t^h	ʧ	6	nigh**t**eouce(8:0)
17	bt^h	bɛt^h	5	oske**bt**ed(4:0)
18	k^hw	k^h	5	le**q**uesh(4:0)
19	nj	n	5	**pn**eudob(3:0)
20	sk^h	s	5	ko**sc**le(5:1)
21	k^h	ʧ	4	ba**ch**t(2:0)
22	t^h	t^hj	4	nigh**t**eouce(8:0)
23	k^hw	k^hj	4	le**q**uesh(4:0)
24	z	ts	4	sa**zz**(5:0)
25	ɹ	ɹh	4	**rh**ypo(3:0)
26	k^h	x	4	ba**ch**t(2:0)
27	s	p^hs	4	**ps**ymno(6:0)
28	ʌf	aʊ	4	zr**ough**(5:0)
29	p^hs	z	3	gor**ps**(6:0)
30	g	gg	3	la**gg**ato(6:0)
31	m	mb	3	opla**mb**(2:1)
32	b	bb	3	ri**bb**ir(8:0)
33	ʒ	zj	3	dei**z**ure(5:0)
34	p^h s	s	3	gor**ps**(6:0)
35	ʤɹ	d	3	wor**dr**abo(3:0)
36	p^h	p^hp^h	3	nu**pp**ola(5:0)
37	bt^h	bɪt^h	2	oske**bt**ed(4:0)
38	ʤɹ	dəɹ	2	wor**dr**abo(3:0)
39	m	mɪn	2	ma**mn**(7:0)
40	ʧ	ʃ	2	ip**ch**arl(4:0)
41	n	ŋ	2	calei**gn**(3:0)
42	ŋ	ng	2	ci**ng**(3:0), jaso**ng**(3:0)
43	ʌf	oʊ	2	zr**ough**(5:1)
44	ʌf	u:f	2	zr**ough**(5:0)
45	p^hs	p^hz	1	gor**ps**(6:0)
46	g	gh	1	a**gh**ach(3:1)
47	g	ʤ	1	mi**gg**est(3:1)
48	ʤɹ	ʤɹj	1	**dr**umotic(2:0)
49	ŋ	n	1	jaso**ng**(3:0)
50	ŋ	g	1	jaso**ng**(3:0)
51	n	nj	1	nu**nn**y(3:0)
52	nj	pn	1	**pn**eudob(3:0)
53	ʤ	g	1	**g**enroi(3:0)
54	ʌf	ɑːugh	1	zr**ough**(5:0)

Table 4.2: LLMs' mistakes regarding consonants. This table includes mistakes with count < 7. The last column shows the pseudo-word(s) where the mistake happens with the part of interest in bold, and the numbers in the parentheses are "count of the golden prediction" : "count of the mistake" among human participants.

real words and the results are presented in Table 4.3.

We tested not only the uncommon mappings, but also the corresponding common mappings (marked with $*$) for comparison. Note that only the specified mistakes are counted. For example, if *th* is wrongly predicted as [ð], it is counted in the first row; but if it is wrongly predicted as [d], it is not counted. This way, we can tell that both LLMs have little problem if the correct mapping is a common one, but err a lot on an uncommon mapping to fricatives. The baseline performance on *th* → [ð], *s* or *ss* → [ʃ], and *s* → /ʒ/ suggests that mistakes are to a great extend avoidable simply by taking the succeeding vowels into account, but LLMs somehow fail to do so.

Correct	**Mistake**	**Total**	**Llama3**	**Mistral**	**baseline**
*th → [θ]	[ð]	216	4	5	37
th → [ð]	[θ]	286	111	107	43
*ss → [s]	[z]	25	1	0	0
ss → [z]	[s]	24	20	12	16
*s → [s]	[z]	44	1	6	5
s → [z]	[s]	30	6	3	4
c → [ʃ]	[s]	190	63	27	47
s → [ʒ]	[z]	5	5	5	5
s or ss → [ʃ]	[s] or [z]	67	42	35	13
s → [ʒ]	[s] or [z]	201	56	64	3

Table 4.3: Validation results based on real English words showing that LLMs' struggle over "uncommon" mappings to fricative consonants. The rows with $*$ are about "common" mappings; they are included to show the asymmetry of model mistakes. Note that only the specified mistakes are counted.

4.1.2 Failure to Handle Silent Consonants

A similar tendency is suggested by rows 2, 3, 4, 5, 8, and 10 in Table 4.1 (also rows 15, 27, and 31 in Table 4.2). When consonant letter should be silent from the human perspective, LLMs can often make an alternative decision, which is again based on the most common mapping from that consonant letter. We collected real words with silent letters *p, b, g, k, m(n), (r)h, w(r)* for the validation experiment, and the result is presented in Table 4.4.

Except for the silent *h* in *rh*, both LLMs exhibited the tendency to sound the letter to a various extent. When it comes to *h* in *rh* → -, not only LLMs prevail, the baseline also avoids adding an [h] sound.

Correct	Mistake	Total	Llama3	Mistral	baseline
p → -	[p^{h}] or [p]	67	8	12	12
b → -	[b]	97	41	62	41
g → -	[g] or [ɢ]	116	16	23	41
k → -	[k^{h}] or [k]	61	9	7	12
n in mn → -	[n]	17	3	6	12
h in rh → -	[h]	52	0	2	0
w in wr → -	[w]	147	6	29	6

Table 4.4: Validation results based on real English words showing that LLMs' struggle over silent consonant letters. Note that only the specified mistakes are counted.

4.1.3 Extra [j] Mapped from 'u'

We discuss the semi-vowel [j], which is after all a consonant, in the this section. This sound often appears when there is a vowel letter *u*[3]. Relevant mistakes are not only in rows 15, 19, 22, and 23 in Table 4.2, but, more importantly, rows 2 and 5 in the forthcoming Table 4.6[4]. The lesson seems to be: LLMs simply mess things up when it comes to [±j]. Sometimes, an extra [j] is added before [uː]; on other occasions, [j], which is given by human participants, is missing from the model prediction.

Our validation data includes three groups of real words. Apart from *u* → [uː] and *u* → [juː], we also collected words with where *qu* is mapped to [k^{h}w] or [kw]. Note that there is no vowel sound for the *qu* part, while the vowel letter that immediately follows is mapped to a vowel sound. This vowel sound, is never [uː] as it conflicts the preceding semi-vowel [w]. For example, *quest* would satisfy while *queue* is not collected in the group. The reason behind this group is that we find LLMs less often add [j] on such pseudo-words as *lequesh* and *vasquido*. While both [+j] and [−j] are *in theory possible* in pseudo-words like *minudo* and *dumy*, this is not the case with *lequesh* and *vasquido*.

It turns out that, on the validation dataset, both LLMs show strong tendency to add a redundant [j] upon seeing *u*, and the baseline avoids doing this by taking the preceding consonant letter into consideration (Table 4.5). On the other hand, however, LLMs are less likely to leave out a required [j]. Put together, these suggest an overall [+j] tendency, which partly aligns with our finding on pseudo-words. Also, LLMs do perform better on the *qu* → [k^{h}w] or [kw] case, adding [j] very few times.

[3]We avoid saying [j] is "mapped" from *u* as it is considered as attached to the preceding consonant sound, such as m → [mj] in *mule*.

[4]Some /j/-related issues are reflected in the table for vowels due to how we transcribe and log the predictions: If a there is a semi-vowel (i.e. [j] or [w]), always include it, whether our main concern is a consonant sound or a vowel sound.

Correct	Mistake	Total	Llama3	Mistral	baseline
qu → [k^{h}w] or [kw]	+[j]	282	5	1	1
u → [uː]	+[j]	119	26	38	9
u → [juː]	-[j]	132	9	5	10

Table 4.5: Validation results based on real English words showing that LLMs' tendency to incorrectly add the semi-vowel [j] when mapped from *u* but not when mapped from *qu*. Note that only the specified mistakes are counted.

4.2 Tendencies with Vowels

Analysing 'mistakes regarding vowels proves harder because: 1) for a given POI, vowel predictions can vary in length[5], 2) on 5 pseudo-words, namely *sejor*, *duoy*, *noge*, *gly* and *molia*, there is more than one golden prediction, making the mistake hard to formulate, and 3) in general, vowel mistakes take more forms than consonant mistakes do and one can hardly read anything into Tables 4.6, 4.7, and 4.8.

To address 2), we simply exclude the 5 aforesaid pseudo-words from this part. To address 1) and 3), we divide all model mistakes into two broad categories: aligned mistakes (where the model prediction and the golden prediction have the same length, namely 1 vs 1 or 2 vs 2[6]; note that "ː" does not count as a standalone character) and misaligned mistakes. Aligned mistakes are further divided into 1 vs 1 cases, which are basically monophthongs, and 2 vs 2 cases, which we refer to as "diphthongs" loosely because [iːə], for example, is technically not a diphthong and considered as two syllables.

This way, we can take more holistic approaches and examine the mistakes regarding monophthongs by looking at how the sound shifts from the golden prediction to the model prediction (subsection 4.2.1.), examine the mistakes regarding diphthongs by looking at the glide pattern (subsection 4.2.2.), and examine the general tendency of misaligned mistakes (subsection 4.2.3.). After that, we discuss three specific tendencies regarding vowels (subsections 4.2.4-6.).

[5]This does happen with consonants as well, such as the silent consonant issue and the /±j/ issue that we just discussed, but they can easily be classified as such.

[6]After excluding few mistakes involving non-vowel sounds such as /uː/ vs /ʊwɛ/ in *jewelim*, we are left predictions with either length 1 or 2.

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
1	ɑː	æ	42	gn**a**lf(5:0), n**a**l(3:0)
2	juː	uː	32	min**u**do(2:0), t**u**ce(2:1), b**u**cry(4:0), b**ui**tal(2:1)
3	uː	ʌ	30	d**u**my(4:0), b**u**sitish(3:2)
4	ə	ɪ	26	san**i**my(5:2)
5	uː	juː	21	b**u**sitish(3:0), r**u**ma(4:0), d**u**my(4:0)
6	ɜː	ɜ	20	b**i**rp(6:0), m**i**rty(5:0), n**u**rdes(5:0), s**u**rr(4:0), sb**u**ry(4:0)
7	oʊ	ɒ	19	t**o**lf(4:0), f**o**cipay(2:1), b**o**sal(5:2), kn**ow**lism(2:0)
8	ɜː	ə	19	sb**u**ry(4:0), b**i**rp(6:0)
9	ɛ	ɑː	19	resc**ar**ce(2:1)
10	iːoʊ	iːə	18	p**eo**tal(2:0)
11	ɛ	iː	13	th**e**mi(2:0), br**ie**nd(3:0)
12	-	ɛ	13	cath**e**krod(4:0)
13	ɒ	oʊ	13	**o**nooc(4:0), d**o**bd(6:0)
14	uː	uːə	13	j**ewe**lim(2:0)
15	ɑː	ɑ	13	gn**a**lf(5:0), r**a**rsh(5:0), at**a**rs(4:0)
16	ɒ	ə	13	**o**nooc(4:1)
17	-	ə	12	cath**e**krod(4:1), pod**e**(3:0)
18	ɒ	ɑ	12	**o**nooc(4:0), d**o**bd(6:0), d**o**ppy(3:0), d**o**kt(6:0)
19	uːi	aɪ	12	t**u**y(2:0)
20	ɪ	ɛə	12	sm**ei**r(5:0)
21	aɪ	ɪ	11	h**i**cent(2:0), t**y**hoch(2:0)
22	juː	juːɪ	11	b**ui**tal(2:0)
23	iː	ɛ	11	wallp**e**te(7:0), v**e**de(2:1)
24	ɪ	aɪ	10	m**y**mn(2:0), dokn**i**tte(4:1)
25	eɪ	æ	10	s**a**ze(6:0), fl**ai**t(5:0)
26	ɪ	ə	9	matol**y**mb(3:0), vak**i**r(2:0)
27	ɛ	aɪ	9	br**ie**nd(3:0)
28	eɪ	aɪ	9	s**ai**p(3:2), fl**ai**t(5:0)
29	ə	ɒ	7	**a**mmock(3:1)
30	juː	ʌ	7	b**u**cry(4:2)
31	ɔː, ə	ɔ	7	sej**o**r(2:0)
32	ɔɪ, uːi	uːɔɪ	7	d**uoy**(2:0)

Table 4.6: LLMs' mistakes regarding vowels. This table includes mistakes with count ≥ 7. The last column shows the pseudo-word(s) where the mistake happens with the part of interest in bold, and the numbers in the parentheses are "count of the golden prediction" : "count of the mistake" among human participants.

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
33	ə	ʊə	2	osk**u**r(2:0)
34	ɜː	ʊ	2	s**u**rr(4:0), sb**u**ry(4:0)
35	iːə	juː	2	tas**eu**m(4:0)
36	uː	uːɛ	2	j**ewe**lim(2:0)
37	oʊ	ə	2	b**o**sal(5:0), f**o**cipay(2:0)
38	ɜː	ɪə	2	m**i**rty(5:0)
39	ə	aɪ	2	san**i**my(5:0)
40	eɪ	ɑː	2	s**a**ze(6:0)
41	ɛ	ɪə	2	br**ie**nd(3:0)
42	iːə	uː	2	tas**eu**m(4:1)
43	uːi	ui	2	t**uy**(2:0)
44	ɛ	ə	2	th**e**mi(2:0), **e**llo(4:0)
45	ʌ	ə	2	k**u**n(3:0), d**u**gnore(3:0)
46	oʊ	aː	2	t**o**lf(4:0)
47	ɛ	ɪ	2	**e**ngtant(4:1)
48	ɛ	ɑ	2	resc**a**rce(2:0)
49	ɪ	eɪ	2	sm**ei**r(5:0)
50	ɒ	ɔ	2	d**o**bd(6:0)
51	ə	ɑ	2	**a**mmock(3:0)
52	uː	u	2	r**u**ma(4:0)
53	aɪ	ʊɪ	2	g**ye**(2:0)
54	uː	ʊwɛ	2	j**ewe**lim(2:0)

Table 4.7: LLMs' mistakes regarding vowels. This table, as well as the next one, includes mistakes with count < 7. The last column shows the pseudo-word(s) where the mistake happens with the part of interest in bold, and the numbers in the parentheses are "count of the golden prediction" : "count of the mistake" among human participants. Note that some pseudo-words have two golden predictions, and the first number in the parentheses is the count of each.

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
55	i	iː	1	bety(2:0)
56	ɜː	ʊə	1	s**u**rr(4:0)
57	oʊ	aʊ	1	kn**ow**lism(2:0)
58	iːə	jʊ	1	tas**eu**m(4:0)
59	uː	uwɛ	1	j**ewe**lim(2:0)
60	ɑː	a	1	gn**a**lf(5:0)
61	ɒ, oʊ	ɔː	1	n**o**ge(2:0)
62	uː, oʊ	ɔ	1	m**o**lia(2:0)
63	ɜː	ɪ	1	m**i**rty(5:0)
64	weɪ	wɛ	1	beq**uay**(5:0)
65	oʊ	ou	1	kn**ow**lism(2:0)
66	ɔɪ, uːi	juːoɪ	1	d**uoy**(2:0)
67	iːə	juːə	1	tas**eu**m(4:0)
68	ɪ	i	1	n**i**k(5:0)
69	ɒ, oʊ	ə	1	n**o**ge(2:0)
70	ɑː	aː	1	at**a**rs(4:0)
71	ə	ɜ	1	osk**u**r(2:0)
72	juː	ɜː	1	b**u**cry(4:0)
73	ə	ɛ	1	ind**e**r(2:0)
74	iː, aɪ	ɪ	1	gl**y**(2:0)
75	juː	ɪ	1	min**u**do(2:0)
76	juː	ʊ	1	min**u**do(2:0)
77	eɪ	ɛ	1	fl**ai**t(5:0)
78	ɔɪ, uːi	uː	1	d**uoy**(2:0)
79	ɔɪ, uːi	uoɪ	1	d**uoy**(2:0)
80	ɔɪ, uːi	juːɔɪ	1	d**uoy**(2:0)
81	ɔɪ, uːi	uɔɪ	1	d**uoy**(2:0)
82	uː	uːwə	1	j**ewe**lim(2:0)
83	oʊ	ɑ	1	f**o**cipay(2:0)
84	weɪ	iː	1	beq**uay**(5:0)
85	ɜː	jʊə	1	sb**u**ry(4:0)
86	ɜː	juː	1	sb**u**ry(4:0)
87	ɪ	ɪə	1	vak**i**r(2:0)
88	uː, oʊ	ə	1	m**o**lia(2:0)
89	iː	eɪ	1	v**e**de(2:0)
90	juː	u	1	t**u**ce(2:0)
91	juː	uːɪ	1	b**ui**tal(2:1)
92	iːoʊ	ɪoʊ	1	p**eo**tal(2:0)
93	eɪ	ɛɪ	1	fl**ai**t(5:0)
94	ɔɪ, uːi	ʊɔɪ	1	d**uoy**(2:0)
95	iːə	ɪə	1	tas**eu**m(4:0)
96	uː	ʊə	1	j**ewe**lim(2:0)
97	uː	ʊɛ	1	j**ewe**lim(2:0)
98	ɒ	uː	1	**o**nooc(4:0)
99	uːi	juːi	1	t**uy**(2:0)
100	uːi	waɪ	1	t**uy**(2:0)

Table 4.8: Continued from the last table.

4.2.1 Vertical Shifting of Monophthongs

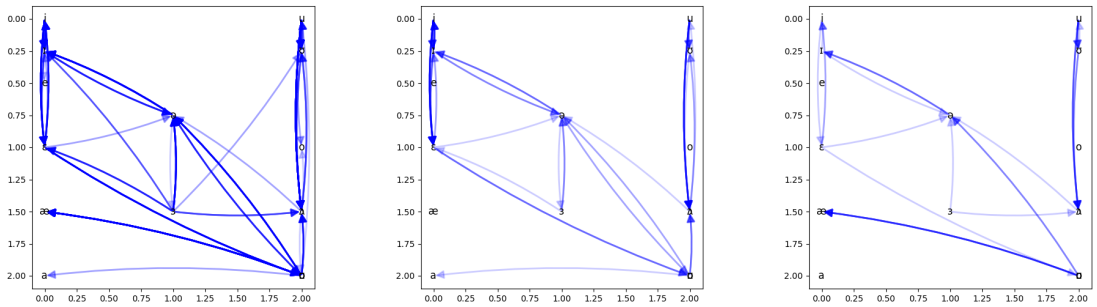
When the golden prediction and the incorrect model prediction are both monophthongs, the difference can come in three features: length (±ː), rounded-ness, and position. Change in length and round-ness, however, are usually intertwined with the shift in position. That is, if a model or a human predicts [iː] rather than [ɒ], for example, that is usually because that *position* is chosen by the model/human given the spelling of the pseudo-word, not because the model/human decides the sound should be long or unrounded[7]. We can only confidently say there is a tendency in changing the length or rounded-ness upon seeing such changes *in the same position*, like [ɑː] vs [ɒ]. Such tendency, however, is not found in Tables 4.6, 4.7, and 4.8. Consequently, we focus on the positional shift in the current study.

We plot the monophthong shifts for all 7 LLMs (Figure 4.1). To do this, we first define the coordinates for all English monophthongs:

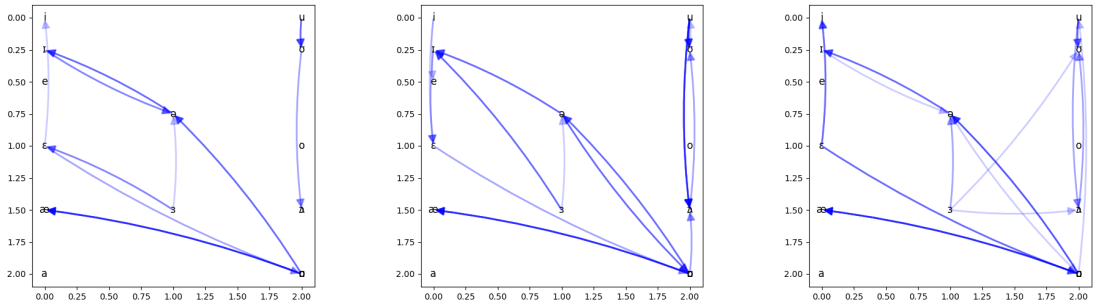
/i/: (0, 0) - the most front-high point
/ɪ/: (0, 0.25)
/e/: (0, 0.5)
/ɛ/: (0, 1)
/æ/: (0, 1.5)
/a/: (0, 2)
/ə/: (1, 0.75) - the central-mid vowel but not (1, 1)
/ɜ/: (1, 1.5)
/u/: (2, 0)
/ʊ/: (2, 0.25)
/o/: (2, 1)
/ʌ/: (2, 1.5)
/ɔ/: (2, 1.5)
/ɑ/: (2,2)
/ɒ/: (2,2) - the most back-low point.

Note that the front vowels are defined to have the same x-coordinates. In fact, the tongue position is always more back when lowered, therefore the meaning of the x-coordinates should not be understood as the exact positions. Similarly, when we say "vertical shifts" later, we simply mean front/central/back vowels are still front/central/back vowels after the shift. Also, /ɪ/ and /ʊ/ are a bit more central than other front and back vowels, but in our practice they are defined to have x-coordinates 0 and 2 respectively.

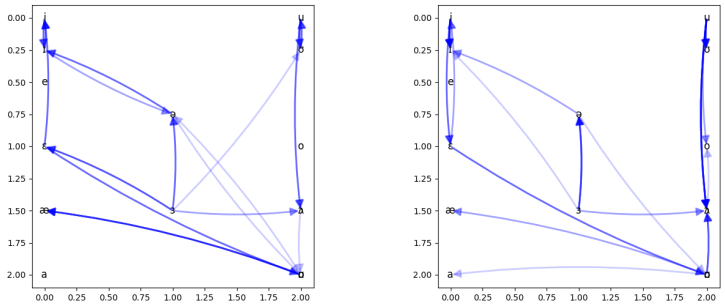
[7] Actually, there are only two pairs of English vowels that sit the same position, namely ([ʌ], [ɔː]) and ([ɑ], [ɒ]). That means position is in some sense correlated with length and rounded-ness.



(a) overall (b) Qwen2 (c) Gemini2



(d) Gemini3 (e) GPT4 (f) GPT5



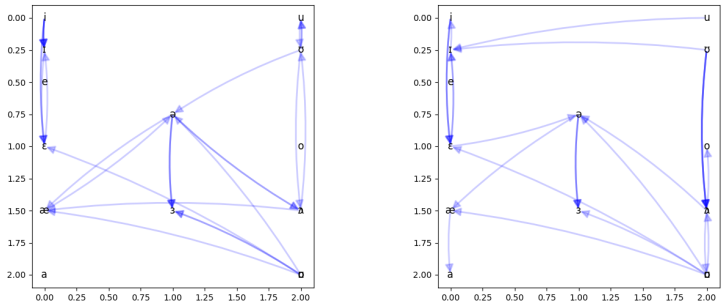
(g) Llama3 (h) Mistral

Figure 4.1: Monophthongs shifting from the golden prediction on pseudo-words. The less transparent the arrow is, the more frequent the way of shifting happens.

Now the plots in Figure 4.1 suggest that vertical shifts are a major type of shifts. At the first glance, this seems unsurprising, for it is more reasonable to give a wrong prediction that is close to the golden prediction rather than one that is far off and sounds very different. Actually, for En-

glish speakers and learners, it is more likely to get confused with vowels on the same vertical line. However, given the fact that all "mistakes" examined here are already based on golden predictions, this is telling the message once again: LLMs are not as good as humans performance-wise, but the way they struggle and err are still human-like.

We further created a **vowel validation dataset** of 150 real English words, 10 for each of the vowel sounds /i/, /ɪ/, /ɛ/, /æ/, /ə/, /u/, /ʊ/, /ʌ/, /ɔː/, /ɒ/, /aʊ/, /aɪ/, /eɪ/, /oʊ/, and /ɔɪ/. For all monophthongs in the dataset, we plotted the aligned mistakes made by Llama3 and Mistral in Figure 4.2 and found the same vertical pattern. Numerically, Table 4.9 tells that LLMs tend to give a wrong prediction through vertical shift, doing this $\geq 60\%$ of the times.



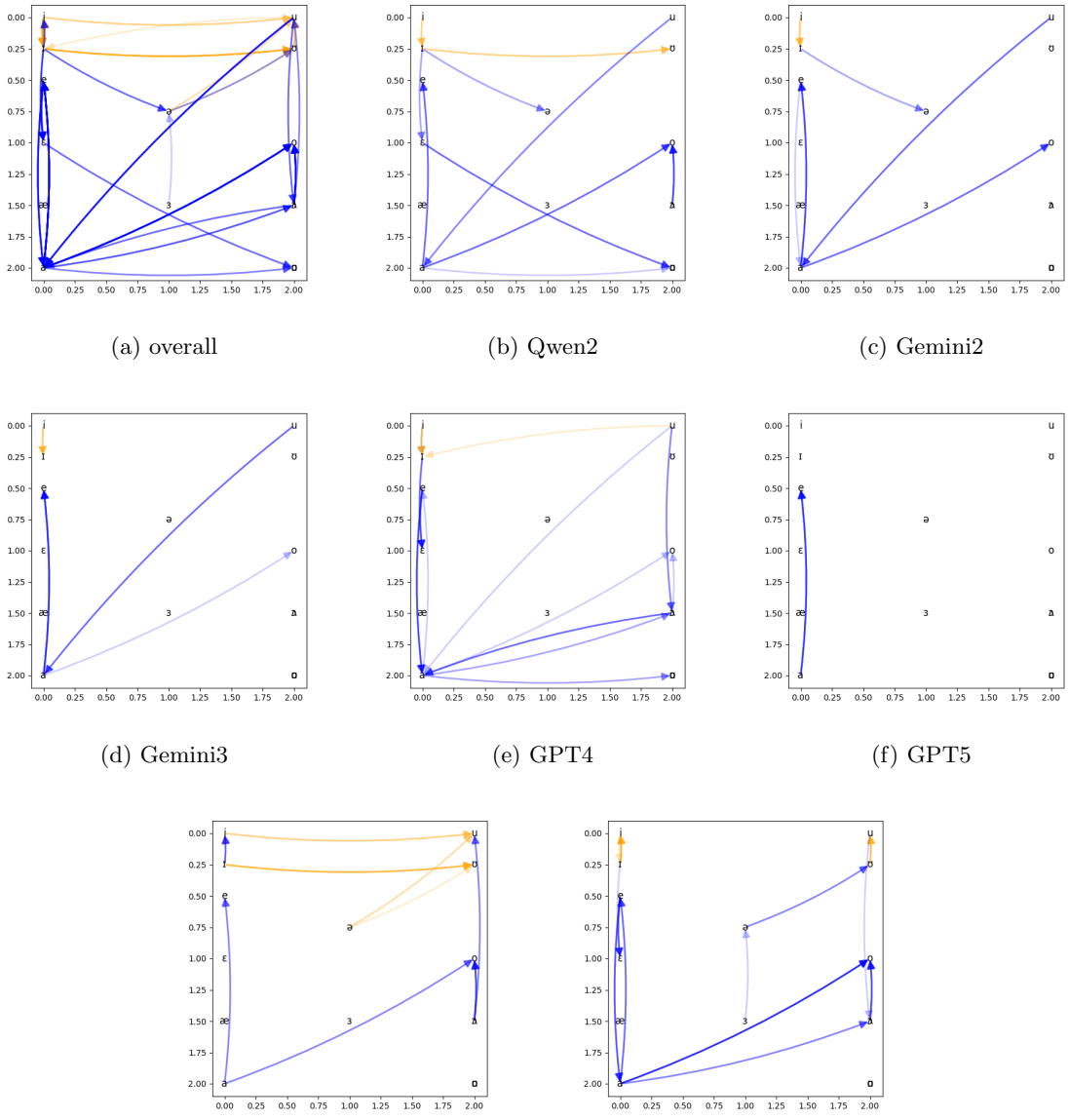
(a) Llama3 (b) Mistral

Figure 4.2: Monophthong shiftings from the golden prediction on the validation dataset (real words). The less transparent the arrow is, the more frequent the way of shifting happens.

We also did the same test for aligned mistakes involving diphthongs, considering the shifts happened between corresponding vowels. For example, if the golden prediction is /aʊ/ and the model prediction is /eɪ/, then two shift are plotted, namely /a/ → /e/ and /ʊ/ → /ɪ/. It turns out that vertical shifts are still the major type of shift even though the mistakes involving diphthongs could be of a different nature (Figures 4.3 and 4.4 and Table 4.9).

	monophthongs	**diphthongs**
Llama3 (validation)	16 out of 27	7 out of 11
Mistral (validation)	18 out of 27	6 out of 10
All 7 LLMs (pseudo-words)	260 out of 426	119 out of 191

Table 4.9: Occurrences of vertical shifting out of all aligned mistakes (monophthong → monophthong or diphthong → diphthong). In the diphthongs column, count is based on the vowel sound - a diphthong contains 2 vowel sounds and thus can contribute up to 2 to the count.

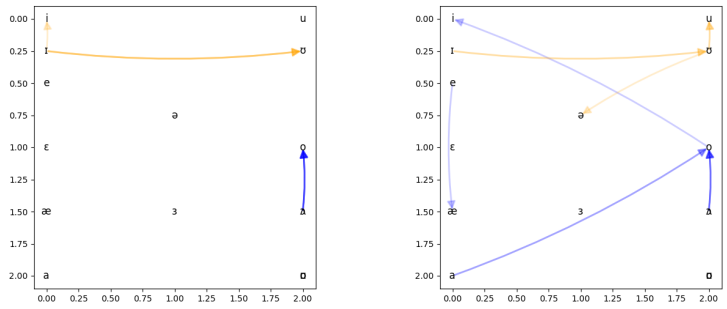


(a) overall (b) Qwen2 (c) Gemini2

(d) Gemini3 (e) GPT4 (f) GPT5

(g) Llama3 (h) Mistral

Figure 4.3: Vowel sounds in diphthongs shifting from the golden prediction on pseudo-words. Blue arrows stand for the beginning vowel in a diphthong and orange arrows stand for the ending vowel. The less transparent the arrow is, the more frequent the way of shifting happens.



(a) Llama3 (b) Mistral

Figure 4.4: Vowel sounds in diphthongs shifting from the golden prediction on the validation dataset (real words). Blue arrows stand for the beginning vowel in a diphthong and orange arrows stand for the ending vowel. The less transparent the arrow is, the more frequent the way of shifting happens.

4.2.2 Shortened Glide of Diphthongs

A diphthong (defined in the aforesaid loose sense) is the process of gliding from one vowel sound to another. Though we still do not know what this indicates, we do see a tendency for LLMs to make shorter glides when we compare the model predictions with the golden ones. The distance of the glide is defined as the Euclidean distance between the starting and ending vowels based on the coordinates given in the previous subsection. For every aligned diphthong mistake, we compute the distances of the model prediction and the golden prediction and then perform a two-tailed t-test. It turns out that for 4 out of the 7 LLMs considered, the model prediction tends to make a glide that is ≥ 1 shorter (Table 4.10). Given that the entire space is 2×2, 1 is a considerable difference. The same test was carried out on the diphthongs in the **vowel validation dataset** and similar result was found for Llama3, but not for Mistral. There could be two reasons: either the count($= 8$) is too small for the tendency to show or Mistral just does not have such tendency as we also fail to find it for Mistral on pseudo-words.

model	Mean Diff.	p	count
Llama3	1.556	**0.000**	13
Mistral	0.438	0.075	31
Qwen2	1.154	**0.000**	18
Gemini2	1.075	**0.001**	15
Gemini3	1.147	**0.000**	13
GPT4	-0.323	0.182	20
GPT5	1.500	**0.000**	7
overall	0.766	**0.000**	117
Llama3 (validation)	0.947	**0.025**	7
Mistral (validation)	0.685	0.118	8

Table 4.10: Comparing the gliding distance of diphthongs with the two-tailed t-test (golden - prediction). The first seven rows show results based on the pseudo-words. The final two rows show results based on the validation dataset (real words).

4.2.3 Influence of String Length

Now we turn to misaligned mistakes, where the length of the model predictions is different from that of the golden prediction. After looking at all the mistakes, we suspect that the LLMs are influenced by the length of the POI in the pseudo-word: if the POI is made of one vowel letter, the model tends to predict one vowel sound; if the POI is made of two vowel letters, the model tends to predict two vowel sounds (e.g. [oʊ] for *t**o**lf* instead of [ɒ] and [ɛə] for *sm**ei**r* instead of [ɪ], as shown in rows 7 and 20 in Table 4.6). To examine whether this tendency exists, we counted mistakes influenced by POI length on both pseudo-words and the **vowel validation dataset**. If a misaligned wrong prediction is the same long as the POI, then it is considered influenced; it can be longer or shorter than the golden prediction. The results are shown in Table 4.11.

model	longer	shorter	count
Llama3	41	5	70
Mistral	27	18	71
Qwen2	30	15	64
Gemini2	45	7	61
Gemini3	42	7	56
GPT4	47	14	71
GPT5	46	7	59
Llama3 (validation)	3	6	13
Mistral (validation)	8	15	24

Table 4.11: Count of cases where a model prediction, not aligned with the golden prediction, is the same long is the part of interest (POI) in the pseudo-word - or real word in the final two rows. "Longer/shorter" in the header means the model prediction is longer/shorter than the golden prediction.

Note that whether the "longer" column has a greater value than the "shorter" column does is

not our concern here; that depends on the dataset structure - for a pseudo-word/real word with a monophthong golden prediction, a misaligned model prediction must be longer. Instead, our concern is: among all misaligned mistakes (the "count" column), how many are influenced by the length of POI (either longer or shorter)? The table tells that model predictions has a $\geq 70\%$ chance, if not higher, to align with the POI in terms of length, suggesting that LLMs are influenced by the POI length.

Row	Golden Pred.	Model Pred.	Count	Pseudo-words
1	ɛɹ	ɛəɹ	10	gh**are**(4:0)
2	ɔːɹ	ɔɹ	6	all**ore**(5:0), f**ore**pinded(6:0)
3	ɪɹ	ɛɹ	6	hudc**ere**(3:0)
4	ɪɹ	ɪəɹi	4	hudc**ere**(3:0)
5	ɔːɹ, wɑːɹ	ɔɪɹ	4	pem**oir**(2:0)
6	ɔːɹ	ɔːɹɛ	4	all**ore**(5:0)
7	ɪɹ	əɹ	4	hudc**ere**(3:0)
8	ɛɹ	ɑːɹ	4	gh**are**(4:0)
9	ɜːɹ	ɜɹ	4	g**yr**tle(3:0)
10	əɹ	jʊə	4	oth**ure**(2:0)
11	ɔːɹ, wɑːɹ	ɔɪəɹ	3	pem**oir**(2:0)
12	ɔːɹ, wɑːɹ	weɪɹ	3	pem**oir**(2:0)
13	ɪɹ	ɪəɹ	3	hudc**ere**(3:0)
14	ɔːɹ	əɹɛ	3	all**ore**(5:0)
15	ɔːɹ, wɑːɹ	wɑɹ	3	pem**oir**(2:0)
16	ɪɹ	ɛ	3	l**ere**(4:0)
17	ɪɹ	ɪə	2	l**ere**(4:0)
18	ɜːɹ	ɜːɹə	2	g**yr**tle(3:0)
19	ɪɹ	əɹi	2	hudc**ere**(3:0)
20	əɹ	ʊɹ	2	oth**ure**(2:0)
21	əɹ	jʊəɹ	2	oth**ure**(2:0)
22	əɹ	jʊɹ	2	oth**ure**(2:0)
23	ɔːɹ	ɔːɪɪ	2	f**ore**pinded(6:0)
24	ɔːɹ	ɔɪɪ	2	f**ore**pinded(6:0)
25	-	ɛə	2	siz**are**(2:0)

Table 4.12: LLMs' mistakes regarding vowels intertwined with the [ɹ] sound. Only mistakes with Count ≥ 1 are included. The last column shows the pseudo-word(s) where the mistake happens with the part of interest in bold, and the numbers in the parentheses are "count of the golden prediction" : "count of the mistake" among human participants. Note that some pseudo-words have two golden predictions, and the first number in the parentheses is the count of each.

4.2.4 Failure to Handle the Silent 'e'

We noticed from the raw data that LLMs more or less fail to handle the silent *e* (Rows 12 and 17 in Table 4.6). The struggle gets worse if the silent *e* is part of *re*. We count relevant mistakes in Table 4.12 (see rows 4, 6, 14, 19, 23 and 24) rather than Tables 4.6, 4.7, and 4.8. Still, the frequency is not high (mostly ≤ 7). Therefore, we use real words to double check if LLMs have the tendency to make a silent *e* sound.

We collect real words for three groups: 1) ending silent *e* following a consonant letter (e.g. *slide*), 2) ending silent *e* following vowel letter + *r* (e.g. *snare* but not *bizarre*), 3) silent *e* sandwiched in two consonant letters (e.g. *boredom*). We count the number of times when the LLM maps the *e* to any vowel, and it turns out that only group 3) seems considerable failure (Table 4.13). When we further created 60 new pseudo-words for "silent *e* in C*e*" and "silent *e* in V*re*" each and tested the two LLMs on them, both started to fail on the "silent *e* in V*re*" group, Llama3 in particular, but not on the "silent *e* in C*e*" group. This make us conclude that LLMs in general possess the knowledge about the silent *e* at the end of a word, but not those in the middle of a word. Also, even if LLMs can predict the word-end *e* in *re* to be silent on real words, they still can be bewildered when it comes to pseudo-words.

Correct	**Mistake**	**Total**	**Llama3**	**Mistral**	**baseline**
e in Ce → -	any [V]	75	3	1	7
e in Vre → -	any [V]	77	4	0	NA
e in CeC → -	any [V]	72	10	26	66
e in Ce → - (new pseudo-words)	any [V]	60	2	0	20
e in Vre → - (new pseudo-words)	any [V]	60	21	8	NA

Table 4.13: Count of cases where the LLM fails to decide *e* is silent. The first three rows show results based on real words. The last two rows show results based on two new batches of pseudo-words. In the first column, C can be any consonant letter (excluding *w, y*) and V can be any vowel letter, namely *a, e, i, o ,u*. In C*e* and V*re*, *e* is the last letter in the word. The last column for *_re* is "NA" because the baseline algorithm happens to map *_re* to [m], which contains no extra vowel but also makes no sense.

4.2.5 Extra [ə] in Rounded Vowels

From Table 4.12, we see a lot of redundant [ə] added in rounded vowels (rows 1, 4, 11, 13, etc.). This is a curious tendency for which we do not yet have a plausible explanation. Take *ghare* as one example. If the golden prediction [ɛɹ] is made by speakers of North American English, then the "corresponding" prediction in British English - or simply the unrounded version - should be [ɛə] rather than [ɛəɹ]. The latter is more like a mixture of both versions.

We are only able to validate the tendency with real words. To formulate: if the golden prediction is [Vɹ], where V can be any vowel other than [ə] (as models seldom add another [ə] into [əɹ]), then the LLM tends to make the prediction [Vəɹ]. We collected real words by this requirement and Table 4.14 shows that such mistake is again frequently made by Llama3 and Mistral, while it is easily avoided by the baseline, which is trained on a single dictionary.

Correct	Mistake	Total	Llama3	Mistral	baseline
Vr → [Vɹ] (where V≠[ə])	[Vəɹ]	49	14	23	3

Table 4.14: Validation results based on real English words showing LLMs' tendency to add a redundant ə in the middle of a rounded vowel. Note that only the specified mistakes are counted. V in V*r* can be any vowel letter (including *y*); V in [Vɹ] and [Vəɹ] can be any vowel sound (excluding [ə]).

4.2.6 Failure to Centralise Unstressed Vowels

Row 4 in Table 4.6 suggests about LLMs' failure to centralise a vowel (→ [ə]) in an unstressed syllable. This is also a common mistake among English learners[8]. Since this issue is tested only once in our pseudo-word list, we want to use real words to double check.

First, we collected 120 words where V → [ə] happens. When a word is given to an LLM, the model get the stressing pattern wrong in the first place, making the centralisation irrelevant. Therefore, we filtered out the words on which at least one of Llama3 and Mistral predicts the stressing pattern wrong. This way, we are left with 74 words and both LLMs know where the unstressed syllable is. The result is presented in Table 4.15. It turns out that, at least on real words, LLMs do not fail to centralise the unstressed vowel that frequently (compared with 26 out of 28 attempts on the pseudo-word *sanimy*) but still have a strong tendency to make this mistake [9].

Correct	Mistake	Total	Llama3	Mistral	baseline
V → [ə]	V≠ [ə]	74	21	10	64

Table 4.15: Validation results based on real English words showing LLMs' failure to centralise unstressed vowels. Note that only the specified mistakes are counted. It is already guaranteed that the model decides on the right stress pattern. V in the first column can be any vowel letter (including *y*); V in the second column can be any vowel sound (excluding [ə]).

4.3 Discussion

Since we are exclusively concerned with where golden predictions elude the LLMs, this chapter can be said to study the non-human-likeness of LLMs. After outlining 9 model tendencies (3 for consonants and 6 for vowels), we find that - if we need one key message - most of them suggest that

[8]Not all vowels in an unstressed syllable are centralised , but this fact makes the matter all the more difficult for language learners.

[9]One thesis committee member raised one concern: unstressed vowels are more robust (distinctive, non-reduced) when words are spoken in isolation. This is legitimate; a better task would be ask the models and participants to produce the word in a flow of utterance. For now, we are safe to say that on the pseudo-word *sanimy*, 5 out of 7 human participants reduced the unstressed vowel to [ə], while LLMs fail to do so on 26 out of 28 runs, suggesting that humans exhibit a stronger tendency to reduce the unstressed vowel even on isolated test pseudo-words

the LLM is an inflexible learner who tends to over-simplify the English G2P problem.

All three tendencies regarding consonants are essentially rigid mapping rules: *th, s, c, ss* are always mapped to the most likely sounds ([θ] and [s]), consonants always sound, and [uː] always has a preceding [j]. The influence of vowel string length can also be viewed this way: if we always predict the most frequent sound for a string of vowel letters, we end up with an IPA string that is as long as the letter string. That might also explain why LLMs seldom make a mistake on *ea* → [iː], *ee* → [iː], and *oo* → [uː] - these are the most common mappings from the given letter strings. Similarly, failure to centralise unstressed vowels and failure to handle silent *e*, especially in the C*e*C environment, are easy to account for if we assume LLMs hold onto some rigid mapping rules.

At the moment, what remain totally beyond us are LLMs' tendencies to make shorter glides in Diphthongs and to add an extra [ə] in rounded vowels (Note that such insertion only makes the IPA string length and the letter string length different.).

Chapter 5
Conclusion and Discussion

5.1 Summary of Contributions

While past studies have shown that LLMs have (unsatisfactory) knowledge of English G2P and pronunciation, we do not know to what extent the models memorised facts about the English vocabulary. Our study instead use pseudo-words for the G2P task in order to evaluate LLMs' actual knowledge. Also, we do not judge the predictions in a binary fashion; PanPhon feature vectors and the editing distance enabled us to better measure the similarity between models and humans. We have shown that LLMs' knowledge does have a good deal of human-likeness, not only because 80% of LLMs' predictions are the same as humans' when there is zero inter-human divergence, also because LLMs' bewilderment (measured by the inter-run distance) correlates with humans'. That is, models and humans are dealing with the G2P task more or less the same way.

We also included SpeechT5 to compare the performance of text-only models and text-speech models. Meanwhile, our text-only models cover a scale from small, old models like MPT-7b-Instruct to state-of-the-art ones like GPT5.4. It turned out that if apart from the three small, old models that are generally known as weaker, all text-only models are more or less the same in terms of human-likeness. On the other hand, SpeechT5 is visibly more human-like on several metrics. This suggests certain bottleneck of text-only models on the G2P task and it seems text-speech models should be counted on if more human-like performance is desired.

When we took a closer look where LLMs fail to agree with humans, we found a couple of tendencies and further validated the results using real words. More than half of the tendencies suggest that LLMs actually oversimplify the matter of G2P, sticking to the most common mappings in the English vocabulary. These tendencies can be useful when we try to improve the performance of LLMs and even text-to-speech models on pronunciation tasks.

5.2 Limitations

Given the experiment design and the way we analyse the tendencies, the knowledge that our study is concerned with is mostly the mapping from a certain chunk of letters to a certain segment of sound (e.g. *p* → [p], *ee* → [iː]). Though these mappings are the building blocks of a word-level G2P task, we should note that local mappings are often intertwined with global decisions such as the stress pattern, especially when it comes to vowel letters/sounds. For a given POI in a given pseudo-word, if the LLM predicts "*a* → [æ]" while all human participants predict "*e* → [ə]", we cannot jump to the conclusion that the model's knowledge about "*a* → [æ]" is wrong; instead,

such misalignment might simply have to do with the different stress patterns the LLM and humans decide on[1]. In this regard, a follow up study is necessary to investigate LLMs' prediction of a pseudo-word's stress pattern, number of syllables, etc..

Another fact we should keep in mind is that LLMs are penalised for using wrong symbols in our study. If an LLM predicts, for example, [ʊː] instead of the golden prediction [uː], it is penalised. In fact, we cannot exclude the possibility that such is a result of unclean training data - human users who are discussing English pronunciation on an internet forum, for example, might not differentiate "ʊ" and "u" and simply use the latter for convenience - and, by a loosened standard, one might still believe the model have made the right prediction. SpeechT5 and human participants, on the other hand, are free from such mistakes as their responses are manually transcribed. We decide not to loosen the standard as our underlying assumption is that LLMs can learn English pronunciation and G2P in the modality of text just like humans and text-speech models can learn those in two modalities, and it is fair to penalise wrong symbols produced by the LLM. While this is more a matter of experimental design rather than a limitation, we are still curious as to what degree of uncleanness is within the training data and how that affects LLMs' knowledge of G2P and pronunciation.

There are a few factors we did not control. The current demographics of native English speakers in our study ("US or Canada born") means we are not concerned with dialectal variation and the possible difference between age groups. Admittedly, both are worth taking into consideration given more time and easiness to recruit participants. Another factor has to do with our pseudo-words. Some pseudo-words are not created based on analogy, yet they accidentally contain real English words (*dug* in *d**u**gnore* and *beam* in *deb**ea**mal*, for example) while some do not (like *c**ou**sh* and *sc**oo**dy*). For the first type of word, whether the human/model prediction is biased by the contained real word is, in its own right, a research question to answer. A closely related question is: To what extent are humans and models affected by the analogy between the base word and the pseudo word (*licc**ough*** and *hiccough*, for example)? We leave these questions to follow-up research.[2]

5.3 Future Directions

5.3.1 Tendency in Text-to-Speech Models

Our study shows that SpeechT5, representing the text-to-speech (TTS) models, is in general more human-like than LLMs, while the latter are susceptible to a few tendencies. We are not yet sure if these tendencies are the very reason why SpeechT5 prevails. A study focused on TTS models - not only monolingual ones like SpeechT5, but also multilingual ones like Fish Speech [36] and Whisper (see [49]) - can further tell if they have the same set of tendencies or some other mistakes to make. The methodology would be mostly the same, but validation on in-vocabulary words is impractical as all aforementioned models have no difficulty reading real English. That means we need a much larger test dataset made of pseudo-words. Nonetheless, such exploration is worthwhile as it paves the way to improving TTS models for real-life applications such as pronouncing unseen words or

[1] We try to avoid implying that the local mapping is a *result* of the stress pattern as it is dangerous to make any assumption about humans'/models' cognitive strategy without evidence.

[2] All three issues discussed in this paragraph are raised by the thesis committee members.

proper names and reading text that involves code switching.

5.3.2 Improving Tokenization

We have always suspected that one fact that hurts LLMs' prediction is the tokenization that is unnatural from a human perspective. Llama3's tokenization of *cheebe*, for example, is *che-ebe*, which is unlikely how human participants analyse the string in their mind. This single fact does not lead to the conclusion that LLMs are inherently handicapped in terms of leaning about G2P and pronunciation as in theory it is possible that useful information are still preserved under the overall "unnatural" tokenization scheme during training. However, it is shown by [35] that LLMs can better process the sound of a real English word when its tokenization is better aligned with syllabification, suggesting that LLMs' struggle on G2P tasks might just be a sophisticated type of the "strawberry problem" [14]. Therefore, we are curious to see if a model with an alternative tokenizer can perform better on our task. As it is impractical to pre-train an LLM using the new tokenization from scratch, we are thinking of using the technique introduced in [3]. This technique enables a monolingual model to learn new token embeddings (on a specific classification task) zero-shot. That means it is possible to let the LLM work with a new tokenizer with minimal effort, allowing us to examine the effect of tokenization on the G2P task.

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon Series of Open Language Models, 2023. arXiv preprint arXiv:2311.16867.

[2] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing, 2022. arXiv preprint arXiv:2110.07205.

[3] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4623–4637. Association for Computational Linguistics, 2020.

[4] Abhinaba Basu and Pavan Chakraborty. Measuring and curing reasoning rigidity: from decorative chain-of-thought to genuine faithfulness, 2026. arXiv preprint arxiv:2603.22816.

[5] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do Neural Machine Translation Models Learn about Morphology? In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[6] Jean Berko. The Child's Learning of English Morphology. *WORD*, 14:150–177, 1958.

[7] Stergios Chatzikyriakidis and Anastasia Natsina. LLMs Got Rhythm? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation, 2026. arXiv preprint arXiv:2601.09631.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. arXiv preprint arXiv:2110.14168.

[9] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and Measuring the Geometry of BERT, 2019. arXiv preprint arXiv:1906.02715.

[10] Anthony G Cohn and Jose Hernandez-Orallo. Dialectical Language Model Evaluation: An Initial Appraisal of the Commonsense Spatial Reasoning Abilities of LLMs, 2023. arXiv preprint arXiv:2304.11164.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and Inderjit Dhillon et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, 2025. arXiv preprint arXiv:2507.06261.

[12] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What You Can Cram into a Single Vector: Probing Sentence Embeddings for Linguistic Properties, 2018. arXiv preprint arXiv:1805.01070.

[13] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM, 2023. arXiv preprint arXiv:312.05621v2.

[14] Adrian Cosma, Stefan Ruseti, Emilian Radoi, and Mihai Dascalu. The Strawberry Problem: Emergence of Character-level Understanding in Tokenized Language Models, 2025. arXiv preprint arXiv:2505.14172.

[15] R. I. Damper and J. F. G. Eastmond. Pronunciation by Analogy: Impact of Implementational Choices on Performance. *Language and Speech*, 40:1–23, 1997.

[16] Ranul Dayarathne, Uvini Ranaweera, and Upeksha Ganegoda. Comparing the Performance of LLMs in RAG-Based Question-Answering: A Case Study in Computer Science Literature. *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, page 387–403, 2025.

[17] Andrea Gregor de Varda, Daniele Gatti, Marco Marelli, and Fritz Günther. Meaning Beyond Lexicality: Capturing Pseudoword Definitions with Language Models. *Computational Linguistics*, 50(4):1313–1343, 12 2024.

[18] Google DeepMind. Gemini 3.1 Pro Model Card. Technical report, Google DeepMind, 2 2026. Available https://deepmind.google/models/model-cards/gemini-3.1-pro/.

[19] Mirjam Ernestus and R. Harald Baayen. Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch. *Language*, 79(1):5–38, 2003.

[20] Phyllis E. Fischer. *The Sounds and Spelling Patterns of English.* Oxton House, Maine, 1993.

[21] Pisoni DB Frisch SA, Large NR. Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language*, 42(4):481–496, 2011.

[22] Susan E Gathercole and Alan D Baddeley. Evaluation of the Role of Phonological STM in the Development of Vocabulary in Children: A Longitudinal Study. *Journal of Memory and Language*, 28(2):200–213, 1989.

[23] A. C. Gimson. *An Introduction to the Pronunciation of English.* Edward Arnold, London, 1989.

[24] R. J. Glushko. The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5(4):674–691, 1979.

[25] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. The Llama 3 Herd of Models, 2024. arXiv preprint arXiv:2407.21783.

[26] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence, 2024. arXiv preprint arXiv:2401.14196.

[27] Bruce Hayes and Colin Wilson. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry*, 39(3):379–440, 07 2008.

[28] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, 2021. arXiv preprint arXiv:2103.03874.

[29] John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure, 2018. arXiv preprint arXiv:1711.10203.

[31] OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et al. GPT-4o System Card, 2024. arXiv preprint arXiv:2410.21276.

[32] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What Does BERT Learn about the Structure of Language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. arXiv preprint arXiv:2310.06825.

[34] Jinu Lee, Qi Liu, Runzhi Ma, Vincent Han, Ziqi Wang, Heng Ji, and Julia Hockenmaier. Entailment-Preserving First-order Logic Representations in Natural Language Entailment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5729–5742, Vienna, Austria, July 2025. Association for Computational Linguistics.

[35] Disen Liao and Freda Shi. How Tokenization Limits Phonological Knowledge Representation in Language Models and How to Improve Them. In *Tokenization Workshop*, 2026. Available https://openreview.net/forum?id=ziGpKA7tvN.

[36] Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis, 2024. arXiv preprint arXiv:2411.01156.

[37] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for Critique-Correct Reasoning, 2024. arXiv preprint arXiv:2402.14809.

[38] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents, 2025. arXiv preprint arXiv:2308.03688.

[39] Celia Martínez-Tomás, Ana Baciero, Miguel Lázaro, and José A. Hinojosa. What Do Pseudowords Tell us About Word Processing? An Overview. *Frontiers in Language Sciences*, 4, 2025.

[40] Jon Torfi Jonasson Derek Besner Max Coltheart, Eileen Davelaar. *Attention and Performance VI*, chapter 25. Routledge, 1977.

[41] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey, 2025. arXiv preprint arXiv:2402.06196.

[42] David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL, 2016.

[43] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can Language Models Learn to Listen?, 2023. arXiv preprint arXiv:2308.10897.

[44] OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, 3 2026. Available https://openai.com/index/gpt-5-4-thinking-system-card/.

[45] Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019.

[46] Jongseok Park, Kyubyong Kim. g2pE, 2019. Available https://github.com/Kyubyong/g2p.

[47] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. Dissecting Contextual Word Embeddings: Architecture and Representation, 2018. arXiv preprint arXiv:1808.08949.

[48] Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. LLM-Powered Grapheme-to-Phoneme Conversion: Benchmark and Case Study, 2024. arXiv preprint arXiv:2409.08554.

[49] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022. arXiv preprint arXiv:2212.04356.

[50] Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. DocFinQA: A Long-Context Financial Reasoning Dataset, 2025. arXiv preprint arXiv:2401.06915.

[51] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof QA Benchmark, 2023. arXiv preprint arXiv:2311.12022.

[52] Mary Beth Rosson. From SOFA to LOUCH: Lexical contributions to pseudoword pronunciation. *Memory and Cognition*, 11:152–160, 1983.

[53] Robert J. Scholes. *Phonotactic Grammaticality*. Mouton, 1966.

[54] Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. PhonologyBench: Evaluating Phonological Skills of Large Language Models, 2024. arXiv preprint arXiv:2404.02456.

[55] MosaicML NLP Team. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs, 2023. Available www.mosaicml.com/blog/mpt-7b.

[56] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline, 2019. arXiv preprint arXiv:1905.05950.

[57] Richard L. Venezky. *The Structure of English Orthography*. Monton, The Hague, 1970.

[58] Yige Xu, Yongjie Wang, Zizhuo Wu, Kaisong Song, Jun Lin, and Zhiqi Shen. Do Vision-Language Models Truly Perform Vision Reasoning? A Rigorous Study of the Modality Gap, 2026. arXiv preprint arxiv:2604.16256.

[59] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Zhou et al. Qwen2 Technical Report, 2024. arXiv preprint arXiv:2407.10671.

[60] Jinming Zhao, Vineel Pratap, and Michael Auli. Scaling A Simple Approach to Zero-Shot Speech Recognition, 2024. arXiv preprint arXiv:2407.17852.