Gram Projections for Conversation Trajectories: A Polynomial Equivalent of the DCT

Jacob Sussmilch
Heya Enterprises Pty Ltd

Fixed orthogonal transforms are a standard way to turn a variable-length sequence of embeddings into a fixed-size vector: keep the first few transform coefficients. The DCT is the default choice. We introduce the Gram (discrete orthogonal polynomial) projection as its polynomial equivalent for short, aperiodic, drifting sequences such as conversation trajectories, and measure whether the choice matters. The Gram coefficients have *stable semantics*: degree 0 is exactly the mean of the sequence, truncation never changes retained coefficients, degree 1 *is* the least-squares drift, and the drift coefficient reports the same end-to-end drift at every sequence length. The DCT shares the first two properties exactly; the readable coefficients and the length-stable drift belong to the polynomial basis alone. On three conversation-satisfaction corpora at matched dimensionality the choice is then free almost everywhere: truncated spans overlap 94-96%, downstream accuracies tie at the instrument's $\sim\pm$2.5-point resolution, and structural gains (splitting the projection by speaker) transfer across the bases. The exception has a location: on populations that mix very short per-role sequences (one to two turns per role) the polynomial basis is better, by a measured 2.3 points — though the test ran on the corpus that generated the hypothesis, so a content or label confound is not excluded. The mechanism is simple: the DCT's leading coefficient confounds drift magnitude with sequence length. Leaving the DC-plus-smooth class entirely has its own measured price, demonstrated with the DST: zero where trajectory shape adds nothing, 3.5 points exactly where it does. Within the class, choose the basis you can read; the accuracy is free.

1. Introduction

A sequence of d-dimensional embeddings $e_1, \ldots, e_T$ — the turns of a conversation, the sentences of a document, the states of an agent — must often become one fixed-size vector. Mean pooling is the degree-zero answer. The standard refinement keeps several leading coefficients of a fixed orthogonal transform, almost always the type-II Discrete Cosine Transform (the DCT-II, or simply "the DCT"; Ahmed et al., 1974; for sentence encoders, Almarwani et al., 2019): complete, orthogonal, energy-compacting for smooth signals, cheap. Outside NLP, projecting sampled trajectories onto an orthogonal basis and learning on the coefficients is the founding move of functional data analysis (Ramsay and Silverman, 2005); this paper is the discrete-grid instance matched to short turn sequences, with identities the continuous theory does not need to state.

This paper introduces the polynomial member of that family for short aperiodic sequences, then asks whether the choice of transform matters at all. The answer, measured on a three-corpus conversation-satisfaction benchmark (Section 3), is more specific than a tie: the bases tie almost everywhere, and the one place they do not is localized by a length-stratified test. Four results carry the paper:

1. The method (Section 2): Gram projection, discrete orthogonal polynomials on the uniform index grid, with three exact identities: G_0 = mean (mean pooling is recovered, not approximated), degree-independence (an anytime property: keeping fewer coefficients never changes the ones kept), and per-coefficient interpretability (G_1 carries the least-squares drift in a single coefficient; no DCT coefficient is a slope). Together the identities give the coefficients stable semantics: each keeps its meaning as the practitioner's knobs move — truncation depth, and, for the drift coefficient, sequence length (Section 5).
2. Equivalence with the DCT, measured (Section 4): the truncated Gram and DCT subspaces overlap 94-96%; downstream accuracy ties on all three corpora at matched dimensionality (every interleaved-arm CI spans zero at a measured instrument resolution of $\sim\pm$2.5 points); and a structural gain, splitting the projection by speaker, transfers across the bases, confirming it belongs to the projection rather than the polynomial family. The basis with readable coefficients therefore comes at no measured cost.
3. The exception, localized (Section 5): a fourth, degree-1 property predicts where any Gram-DCT gap must live. G_1 reports the same end-to-end drift at every sequence length, while the DCT's leading coefficient varies by up to 1.37x across the corpora's length range (Figure 4), so any gap must concentrate where per-role sequences are shortest and lengths most mixed. A length-stratified test confirms the localization, and two interventional arms bound the mechanism — under uniform truncation to per-role length ≤ 3 the two bases are exactly proportional feature-by-feature, so shortness per se costs nothing and mixed lengths are required (Section 5). The mechanism's causal demonstration remains open.
4. The boundary of the class, with a mechanism (Section 6): the DST (the type-II Discrete Sine Transform, DST-II) lacks a constant basis function yet still captures $\sim$90% of the mean in four harmonics, so it does not collapse to shape-only. It pays a penalty only where trajectory *shape* carries label signal: cost = $\max(0, \text{marginal shape contribution})$, demonstrated across three corpora spanning both signs of that quantity.

The intended takeaway is a decision procedure rather than a horse race. Within the DC-plus-smooth class (the bases whose first component is the constant, so the mean is kept exactly, and whose remaining components are smooth), pick the basis whose coefficients you want to read; for drifting, non-periodic trajectories that is the polynomial one, and on populations dominated by very short per-role exchanges it is also measurably the better one. Outside the class, know your corpus's marginal shape contribution (the accuracy the trajectory's shape adds beyond the mean of its turns; one cheap ablation measures it, Section 6) before you pay the price. The argument compresses to one table:

property	Gram	DCT	DST
degree-0 coefficient is the mean, exactly	yes (§2.1)	yes: $D_0 = G_0$ (§2.2)	no (§6)
anytime truncation: kept coefficients never re-fit	yes (§2.1)	yes (§2.2)	yes (§6)
a drift is one readable coefficient	yes (§2.1)	no: spread over all odd harmonics (§2.2)	no (§6)
drift coefficient stable across sequence lengths	yes (§5)	no: up to 1.37x distortion (§5)	no (§6)
accuracy on the benchmark	ties everywhere; +2.3 where per-role lengths are very short and mixed (§4-5)	ties everywhere but that one place (§4-5)	pays max(0, marginal shape contribution) (§6)

Scope. Everything is offline representation for prediction; the downstream task, corpora (USS-SGD, USS-MWOZ, PRISM), heads, and the seed-ensembled bootstrap calibration are specified in Section

3.

---

2. Gram Projection

2.1 Construction and identities

Let $e_1, \dots, e_T$ in R^d be a sequence of embeddings (L2-normalized turn embeddings in our experiments) and let $\phi_k^{(T)}$ be the degree-k Gram polynomial[1]: the family obtained by Gram-Schmidt orthogonalization of $\{1, i, i^2, \dots\}$ on the uniform grid $\{1, \dots, T\}$ under the unweighted inner product, taken in the classical Gram normalization $\phi_k^{(T)}(1) = 1$. The normalization is part of the definition, not a detail: Gram-Schmidt fixes each polynomial only up to scale, rescaling ϕ_k by c rescales the coefficient below by $1/c$, and this convention pins every ϕ_k to a fixed range at every T — $\phi_0 \equiv 1$, and ϕ_1 falls linearly from $+1$ at the first turn to -1 at the last. That fixed range is what later makes the drift coefficient comparable across sequence lengths (Section 5). The degree-k coefficient is the inner-product projection; equivalently, and identically, it is the degree-k coefficient of the least-squares polynomial fit of the sequence on the grid (orthogonality of the basis makes the projection and the fit one computation):

$$G_k = \frac{\sum_i \phi_k^{(T)}(i)\, e_i}{\sum_i \phi_k^{(T)}(i)^2}, \qquad k = 0, \dots, K,$$

with $K = \min(K_{\max}, T-1)$ and zero-padding for coefficients a short sequence cannot determine. Each G_k is a vector in R^d. Three identities follow from orthogonality on the grid, exactly rather than asymptotically:

- $G_0 =$ mean. ϕ_0 is constant, so G_0 is the arithmetic mean of the sequence: mean pooling is the degree-0 special case of the representation.
- Degree-independence. Coefficients are computed by independent inner products; adding or removing higher degrees never changes lower ones. Truncation is therefore *anytime*: every prefix of the coefficient sequence is the optimal least-squares polynomial summary of its degree.
- Interpretability. G_1 carries the least-squares drift of the sequence in one coefficient: under the normalization above, $-2\,G_1$ is exactly the end-to-end displacement of the least-squares line through the sequence, at every T (the fixed sign and factor come from the ± 1 range of ϕ_1). G_2 is the analogous curvature readout. "The conversation drifted toward X" is a statement about one coefficient.

Figure 1 shows the identities on one dialogue before any of them is used. Each turn is a point, placed by its position along the dialogue's own drift axis, and the degree-0, degree-1, and degree-2

[1] Gram polynomials are also known in the literature as discrete Chebyshev polynomials, and they underlie Savitzky-Golay filtering (Savitzky and Golay, 1964); they are distinct from the continuous Chebyshev polynomials T_k, which are orthogonal under a different weight and on different nodes, and whose identities do not transfer to the uniform grid. We write "Gram" throughout to avoid the collision. The construction itself, a low-order orthogonal-polynomial projection with individually interpretable coefficients, has a continuous-domain precedent in L-moments (Hosking, 1990): the r-th L-moment is the projection of a distribution's quantile function onto the shifted Legendre basis, with degree 0 the mean and higher degrees interpretable shape summaries; the coefficients here are the same construction on a uniform discrete grid, applied to a vector-valued sequence over turn index.

least-squares reconstructions are drawn through the turns. The degree-0 line is the mean: exactly the mean-pooling summary. The degree-1 line adds the drift, and its slope is the single coefficient G_1. The degree-2 curve adds curvature without moving either: all three fits share every coefficient they have in common, which is degree-independence drawn rather than stated.

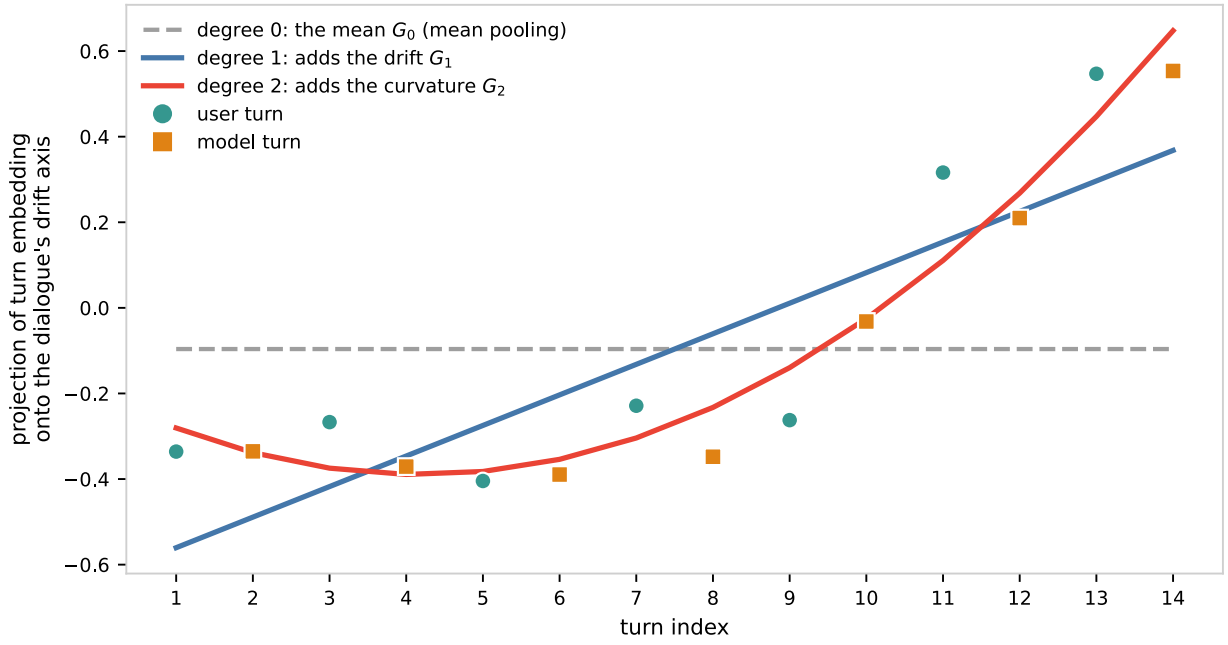

Figure 1: one USS-SGD dialogue's turn embeddings along the dialogue's drift axis (the unit vector along $-2G_1$, the end-to-end displacement of the least-squares line), against turn index; user turns teal circles, model turns orange squares. Overlaid: the degree-0 reconstruction (the mean G_0, exactly mean pooling), degree-1 (adds the drift G_1; the slope is one coefficient), and degree-2 (adds the curvature G_2). All three fits come from one degree-2 projection truncated per curve, so they share every coefficient they have in common; adding a degree never re-fits the ones already kept. The 1-D view is exact for the drawn fits: reconstructions are computed in the full embedding space and projected, and both operations are linear.

The projection is a single $(K+1) \times T$ by $T \times d$ matrix multiply: $O(T \cdot d \cdot K)$, no linear system, and the basis matrix is cacheable per length.

The identities are provable, and also verified numerically on every dialogue in all three corpora; the max absolute error sits at each precision's representational floor (Appendix B).

Degree-independence in practice is an anytime property, and Figure 2 shows what it buys: sweeping the truncation degree $K = 0, \ldots, 6$ of the interleaved projection on USS-SGD, residual energy falls monotonically while downstream accuracy peaks at $K = 0$ and never improves. Description-optimal and prediction-optimal truncation are different knobs, and the coefficients you keep never change as you turn either one. The sweep is the interleaved layout only; the role-split layout (Section 2.3) is evaluated at its standard per-role degree in Section 4.2 (Table 1).

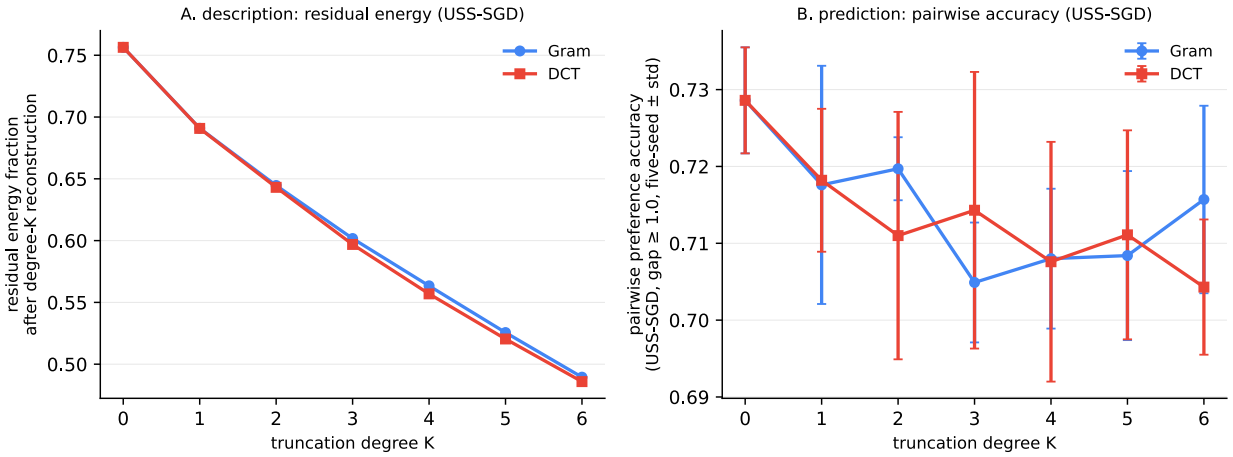

Figure 2: the anytime property in practice (interleaved projection, USS-SGD, both bases, five-seed ± std). Residual energy after the degree-K reconstruction keeps falling through $K = 6$ (0.76 at $K = 0$, 0.60 at $K = 3$, 0.49 at $K = 6$), while pairwise preference accuracy peaks at $K = 0$ (0.729, mean pooling) and wobbles between 0.705 and 0.729 through $K = 6$; on a corpus whose marginal shape contribution is negative (Section 6), added degrees are absorbed, not harmful. Both bases open identically at $K = 0$ ($D_0 = G_0$ = the mean, the shared identity, visible in the data). Truncation depth is a free knob; nothing kept is ever re-fit.

2.2 What the DCT shares

The natural point of comparison is the DCT-II, with basis vectors $\cos\big(\pi(i + \frac{1}{2})k/T\big)$ sampled on the same uniform grid (we write "DCT" for DCT-II throughout). We build it with the *same* least-squares projection we use for Gram: each coefficient is the sequence's projection onto one basis column, normalized by that column's squared norm. Under that shared projection the DCT reproduces the first two identities exactly, not approximately. Its $k = 0$ column is constant, so D_0 equals G_0, the mean, to floating-point precision on every dialogue in all three corpora (the equality is asserted in the code that produced Tables 1 and 3), and its columns are orthogonal on the grid, so its coefficients are degree-independent too. What the DCT does *not* reproduce is the third identity, interpretability: its higher columns are oscillatory cosine harmonics, so an exactly linear drift spreads across all odd harmonics. No single DCT coefficient is a trend, and "the conversation drifted toward X" has no one-coefficient reading; conversely, a slow oscillation is one DCT coefficient but many polynomial degrees. A second asymmetry, the drift coefficient's stability across sequence lengths, is a degree-1 property demonstrated in Section 5. Which prior fits conversation trajectories (short, aperiodic, ending somewhere different from where they began) is the empirical question of Section 4.

2.3 Role-split projection and per-role interpretability

Conversations interleave two speakers. The role-split variant partitions the turns by speaker, projects each speaker's subsequence on its own uniform grid, and concatenates the per-role blocks; role-split(1) keeps per-role degrees 0-1 (per-role mean and drift). By linearity of the projection, the joint-topic midline and the speaker-offset paths are recovered exactly from the per-role blocks: $c_k = (G^u_k + G^m_k)/2$, $d_k = G^u_k - G^m_k$ (verified at machine precision above; the lowercase d_k keeps the speaker offset distinct from the DCT coefficients D_k).

The split is worth stating here because it makes the coefficients' interpretability concrete. Because G_0 is exactly the mean and G_1 exactly the least-squares drift (Section 2.1), each per-role block is a geometric object read directly off the trajectory: a per-role centroid G_0 and a per-role drift G_1.

Figure 3 shows this in a 2-D PCA view of one dialogue's turn embeddings. Read as one interleaved sequence (panel a), the degree-1 fit is a single line drawn through both speakers at once; because the two occupy different regions of the space, it belongs to neither. This is the low-degree aliasing of speaker alternation that a whole-conversation fit cannot avoid. Split by speaker (panel b), each role carries its own readable G_0 (a mean, the star) and G_1 (a drift, the arrow), and the interleaved midline $(G_0^u + G_0^m)/2$ and speaker offset $G_0^u - G_0^m$ are recovered exactly from the two blocks. Nothing is lost, and each coefficient now names one speaker's behavior. This is the per-coefficient interpretability of Section 2.1 in the two-speaker case; whether it also *predicts* better is the separate, downstream question of Section 4.2.

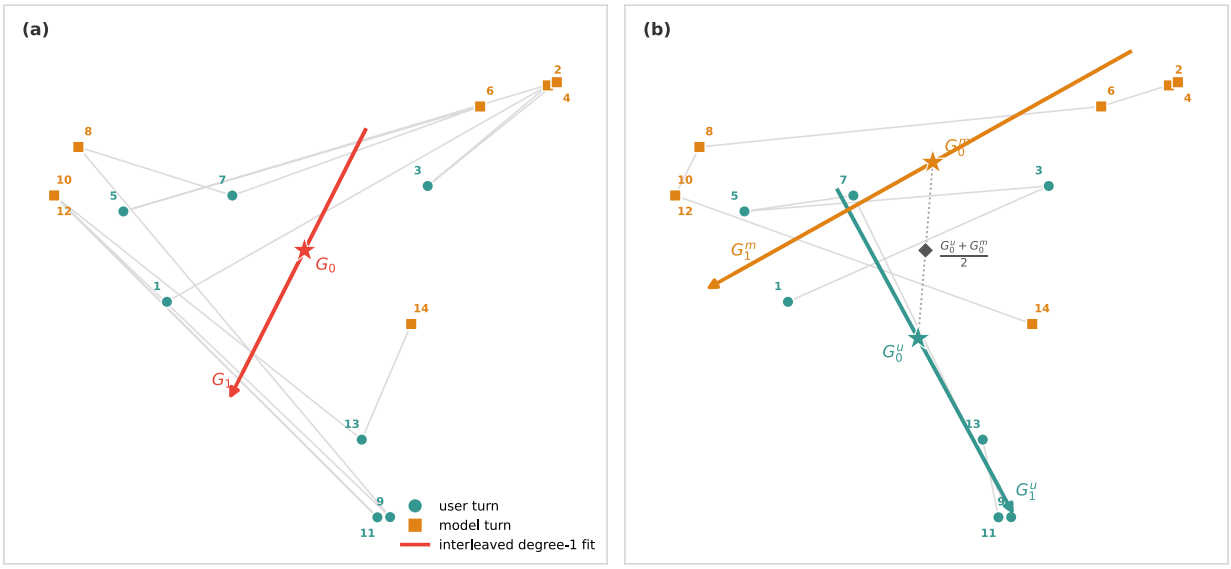

Figure 3: one USS-SGD dialogue's L2-normalized turn embeddings in a 2-D PCA view (user turns teal circles, model turns orange squares, numbered in conversation order; faint grey lines are the raw turn-to-turn paths). The degree-1 Gram coefficients are computed in the full embedding space and projected into the plane, which is exact: the least-squares coefficient is linear in the data and PCA projection is a linear map, so the drawn fit is the true G_1 projected, not a 2-D surrogate. (a) interleaved: the single degree-1 fit, its mean G_0 (star) and drift G_1 (arrowhead), passing between the two speakers' regions, belonging to neither. (b) role-split: each speaker's own G_0 (star) and G_1 (arrowhead), with the recovered midline $(G_0^u + G_0^m)/2$ (grey diamond, halfway between the two stars); every coefficient is a statement about one speaker.

In this paper the split serves exactly one further purpose: a representative *structural* modification whose gains should transfer across bases if they belong to the representation rather than the polynomial family (Section 4.2). Why the split helps, when to prefer it, and what it wins downstream are beyond this paper's scope.

3. Evaluation Setup

One protocol serves every result in the paper; the table specifies it.

setting	choice
task	conversation-level satisfaction as pairwise preference accuracy, over conversation pairs with mean-rating gap >= 1.0
head	`RandomForestRegressor` (300 trees, min_samples_leaf 2; Breiman, 2001) on the representation; 5-fold out-of-fold predictions, user-grouped on PRISM
corpora	USS-SGD and USS-MWOZ (Sun et al., 2021): 1,000 long task-oriented dialogues each over SGD (Rastogi et al., 2020) and MultiWOZ (Budzianowski et al., 2018), third-party labels; PRISM (Kirk et al., 2024): 7,951 short first-person-rated chats (conversations without a helpfulness score or with fewer than two turns are dropped)
turn embeddings	all-MiniLM-L6-v2 (Wang et al., 2020; via sentence-transformers, Reimers and Gurevych, 2019), identical for every arm
inference	seed-ensembled predictions (each dialogue's out-of-fold predictions averaged over five training seeds before scoring) with dialogue-resampling bootstrap CIs
instrument resolution	$\sim\pm$2.5 points, measured: the smallest difference this protocol reliably distinguishes; equivalence statements are estimation-framed and never claimed finer
decision vocabulary	*decided* = 95% CI excludes zero; *null* = CI spans zero at the instrument's resolution; *trend* = consistent point estimate, undecided. Tables carry the numbers; prose uses the words
layouts (both 1,536-d)	interleaved: degrees 0-3 of the single turn sequence, in order; role-split(1): each speaker's subsequence projected separately to per-role degrees 0-1, blocks concatenated (Section 2.3)

The resolution is measured, not assumed. Every table arm below is labelled `basis, layout`; "degrees" always means projection degree, and role-split(k) denotes per-role degrees 0..k.

---

4. Equivalence with the DCT

4.1 Theory: where equivalence must hold

Full-spectrum DCT and Gram coefficient sets are both orthogonal linear images of the same sequence, hence exactly interconvertible; for any head invariant to orthogonal maps, full-set equivalence is a lemma, not an experiment. Truncation breaks the lemma: each low-order coefficient of one basis depends on *all* coefficients of the other, so keeping four of each retains different information. The empirical questions are how different the retained subspaces are, and whether a nonlinear head cares.

Subspace geometry. Over sequence lengths $n = 8, \ldots, 26$ (the USS corpora's interleaved range), the spans of {Gram degrees 0-3} and {DCT coefficients 0-3} overlap 0.94-0.96 (mean squared cosine of principal angles; the largest principal angle is 18-25 degrees); at PRISM's shorter interleaved lengths ($n = 4..7$) the overlap is higher still, 0.97-1.0. The subspaces are close but not identical: close enough to predict a tie, distinct enough that the tie is an empirical result rather than an algebraic one.

4.2 The measured tie

Table 1. Gram and the DCT tie; both gain the same amount from role-split.

basis, layout	USS-SGD	USS-MWOZ	PRISM
Gram, interleaved	0.719	0.774	0.614
DCT, interleaved	0.731	0.769	0.614
Gram, role-split(1)	**0.765**	0.775	**0.642**
DCT, role-split(1)	0.760	**0.780**	0.628

Note: Cells give the seed-ensembled pairwise-preference accuracy at rating-gap >= 1.0; every arm has 1,536 dimensions. Seed-ensembled accuracy (the accuracy of the seed-averaged out-of-fold predictions) runs about a point above the mean of per-seed accuracies. Bold marks the best score in each corpus column, a maximum marker at the table's precision rather than a significance statement.

The basis makes no reliable difference: swapping Gram for the DCT in the interleaved layout moves accuracy by at most 1.2 points, every contrast null, with the sign flipping from corpus to corpus — what a true tie looks like at this instrument's resolution (the contrast table below). The *layout*, by contrast, does matter, and by the same amount in each basis: role-split(1) beats interleaved in every cell of Table 1, and the gain is decided within each basis separately — Gram on SGD and PRISM, the DCT on SGD, and decided Fisher-combined across the three independent corpora. The improvement therefore belongs to the split, a representation-level change that removes speaker aliasing, and not to the polynomial family, exactly as a structural mechanism predicts.

contrast (seed-ensembled)	Δ	95% CI	p
DCT − Gram, interleaved, USS-SGD	+0.012	[-0.005, +0.026]	0.166
DCT − Gram, interleaved, USS-MWOZ	-0.005	[-0.021, +0.009]	0.524
DCT − Gram, interleaved, PRISM	-0.001	[-0.025, +0.026]	0.910
Gram: role-split(1) − interleaved, USS-SGD	+0.045	[+0.018, +0.072]	<0.001
Gram: role-split(1) − interleaved, USS-MWOZ	+0.001	[-0.020, +0.021]	0.992
Gram: role-split(1) − interleaved, PRISM	+0.028	[+0.002, +0.051]	0.032
DCT: role-split(1) − interleaved, USS-SGD	+0.030	[+0.002, +0.058]	0.034
DCT: role-split(1) − interleaved, USS-MWOZ	+0.011	[-0.013, +0.034]	0.368
DCT: role-split(1) − interleaved, PRISM	+0.014	[-0.004, +0.032]	0.118
Gram − DCT, role-split(1), PRISM (the Section 5 exception)	+0.015	[+0.002, +0.027]	0.018
DCT role-split gain, Fisher-combined across the three corpora	—	—	0.042

Note: Fractional accuracy deltas at rating-gap >= 1.0, seed-ensembled, with dialogue-resampling bootstrap CIs; readings in the text use the Section 3 decision vocabulary. The first three rows are the basis tie; the middle six are the within-basis role-split gain; the last two are the Section 5 exception and the cross-basis transfer of the split's gain.

Why per-role degree 1. A single-seed ablation over per-role degrees 0..3 prices the choice (Gram basis, gap >= 1.0): role-split(1) is the smallest configuration never worse than any alternative on any corpus, and every within-split contrast is null at the instrument's resolution; the selection is a

parsimony argument, not a significance claim. The one informative cell: dropping the per-role drift (role-split(0)) costs 4.2 points on USS-MWOZ, the one corpus whose marginal shape contribution is positive (Section 6). Discarding all shape is priced exactly where shape is informative, the same $\max(0, \text{marginal shape contribution})$ logic as the DST boundary, applied to truncation depth. Degrees above 1 buy nothing anywhere. Read jointly with Figure 2, the truncation story is complete: on the interleaved layout added degrees never pay; under the split, degree 1 is load-bearing where shape matters and free elsewhere.

---

5. The Exception, Localized: Mechanism and Length Tests

The tie of Section 4 has one exception; this section reports the observation, the mechanism, and the length tests that fix what may be claimed. The observation: on PRISM's short per-role sequences (median 3 turns per speaker), Gram role-split(1) exceeds its DCT counterpart by 1.5 points — decided (the Section 4.2 contrast table), but one of 21 computed contrasts and below the ±2.5-point band, so treated as exploratory.

The candidate mechanism, demonstrated as a basis property (Figure 4), is cross-length comparability. A Gram G_1 reports the end-to-end drift identically at every sequence length: the same conversation-level drift produces the same coefficient for a 3-turn and a 26-turn speaker, exactly (Figure 4, flat at 1.0), and a pure line is contained in that single coefficient at every length. The cosine basis has no such identity: the DCT's leading coefficient for the same total drift varies by 37% across $T = 3, \ldots, 30$, with the distortion concentrated below $T \approx 6$ and vanishing above $T \approx 25$, so across a population of highly variable short lengths D_1 confounds drift magnitude with length. The demonstrated property bites exactly where per-role sequences are short and variable (PRISM) and vanishes where they are long (USS), matching the observed pattern. The mechanism is real. Whether it *causes* the corpus-level exception is the question the tests below answer: the localization it predicts is confirmed; its causal demonstration is not.

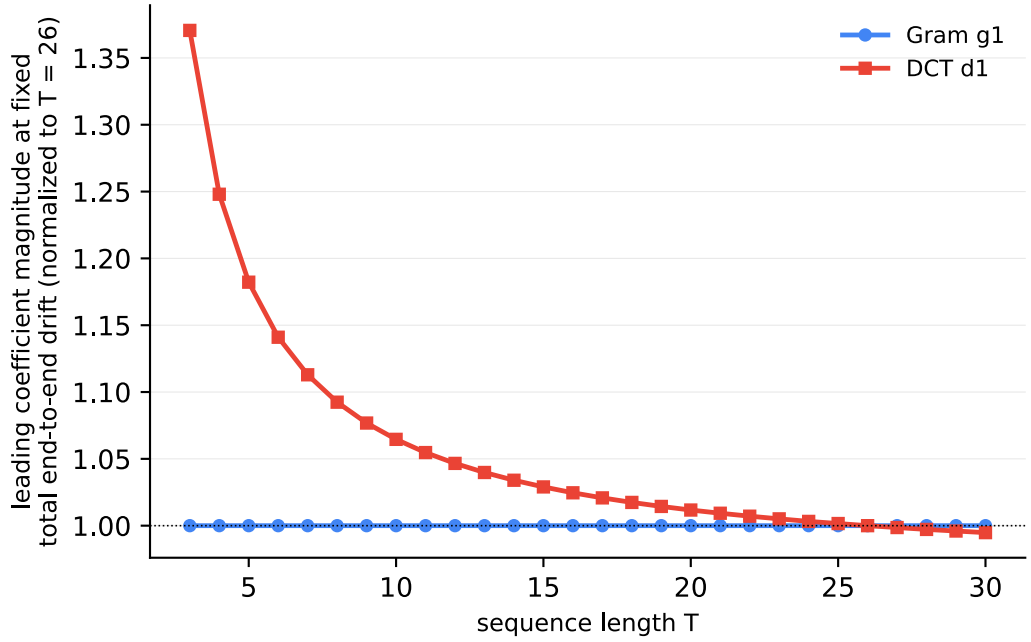

Figure 4: the cross-length property, synthetic and inference-free: leading-coefficient magnitude at fixed total end-to-end drift, normalized to $T = 26$. Gram is exactly flat (the same conversation-level drift produces the same coefficient at every length) while the DCT's leading coefficient for the identical drift swells to 1.37x at $T = 3$, the distortion concentrated below $T \approx 6$ and gone above $T \approx 25$. Demonstrates the mechanism; the corpus-level tests follow.

If the mechanism is right, the deficit must *localize*: concentrated in comparisons involving the shortest per-role sequences, absent where both sides are longer. The test specification has two primary endpoints, a feasibility gate, stratification and fallback rules, and a decision rule for every outcome; the instrument is unchanged from Sections 3-4 (five seeds, seed-ensembled OOF predictions, dialogue-resampling bootstrap at rating-gap >= 1.0). Everything below is reported against those rules, passed or not; the table summarizes the three arms, and the details follow.

arm	design (corpus)	endpoint	result	verdict under the stated rules
E1	observational: pairs stratified by the shorter side's per-role turn count (PRISM)	interaction $I = \Delta_{\text{short}} - \Delta_{\text{long}}$	+2.3 points, CI [+0.3, +4.4], p = 0.022	passes its rule: the gap localizes as predicted
E2	interventional: every role uniformly truncated to depth m (USS)	$J = \Delta(m{=}3) - \Delta(m{=}\text{full})$	-0.4 points, CI [-1.5, +0.7]	null: refutes shortness per se; algebraically blind to the variance form
E3	interventional: per-dialogue depths drawn from $\{2..8\}$ (USS)	Δ_{het} at matched mean depth	+0.6 points, CI [-1.5, +2.6]	inconclusive: under-powered at the stated resolution

E1, observational (PRISM). Pairs of dialogues are stratified by the shorter side's per-role turn count. The mechanism-derived threshold (per-role length <= 5, from the synthetic knee of Figure 4) turned out to be infeasible: PRISM's per-role lengths run min 2 / median 3, with 91% at or below 5, so 99.2% of valid pairs fall short of it. The fallback rule fired instead, a median split at <= 2 turns per role (58% of pairs). The feasibility gate (a placebo contrast between two seed-ensembles of the *same* Gram arm, run through the identical stratified bootstrap) resolved at ±0.9 points, well inside its 3.0-point bar: the test is powered. Table 2 gives the primary result.

Table 2. The exception localized: the Gram-DCT gap lives in the shortest per-role sequences.

quantity (PRISM, seed-ensembled)	Δ (Gram − DCT)	95% CI	p
Δ_{short}: pairs whose shorter dialogue has ≤ 2 turns per role	+0.025	[+0.007, +0.042]	0.008
Δ_{long}: both dialogues ≥ 3 turns per role	+0.002	[-0.011, +0.015]	0.782
$I = \Delta_{\text{short}} - \Delta_{\text{long}}$ (primary endpoint)	+0.023	[+0.003, +0.044]	0.022

Note: Δ is the seed-ensembled pairwise-preference accuracy difference (Gram role-split(1) minus DCT role-split(1)) at rating-gap >= 1.0, evaluated within the pair stratum; joint dialogue-resampling bootstrap (both strata computed from the same resample in every replicate), n_boot = 1000, m-out-of-n at m = 2000, so the CIs are conservative.

The primary endpoint clears its bound — decided (Table 2) — and the decomposition has the prediction's shape exactly: the whole of the exception lives in pairs whose shorter dialogue has at most two turns per role, and vanishes when both sides have three or more (Table 2's two strata). The caveats, as measured. The single-seed mode is directionally consistent but individually null (+1.7 points). The disparity secondary on PRISM (the mechanism's most specific signature: high-disparity pairs should carry more of the deficit than low-disparity pairs) is directionally consistent but null (+1.2 points). The two USS corpora, run as attenuated controls, both failed their power gates (placebo half-widths 3.6 and 3.4 points), so their outcomes are descriptive under the decision rule: MWOZ shows the same signature (+1.8 points nominal), SGD is null, and one wrong-signed descriptive cell (the SGD disparity secondary) is noted, not explained away.

A post-hoc tertile refinement (labeled exploratory in the artifact; added after the primary endpoint was read) completes the natural-length picture: Δ = +2.5 points for pairs at <= 2 turns per role, +0.2 at 3-5, and -1.6 at >= 6 (Figure 5a carries the intervals). That last stratum is only 0.8% of pairs, so its interval is wide, but nothing in it suggests a hidden polynomial advantage at longer

natural lengths. The exception is a <= 2-turns-per-role phenomenon (roughly <= 5 total turns); at mid lengths the bases are equivalent, exactly as Section 4.2 measures.

E2, interventional, uniform truncation (USS). The causal arm truncated every dialogue's roles to a fixed prefix depth $m \in \{2, 3, 4, 6, 8, \text{full}\}$ and rebuilt both role-split(1) representations from the identical truncated subsequences. Its endpoint, $J = \Delta(m{=}3) - \Delta(m{=}\text{full})$ on USS-SGD, came back null (-0.4 points), and the dose-response is null at every level, with a tell: at $m \leq 3$ the CIs collapse to ± 0.01 points. The tell is algebra, not sampling. At per-role length $T = 2$ the DCT's degree-1 basis column is exactly 0.7071 times the Gram column, at $T = 3$ exactly 0.8660 times (proportionality first breaks at $T = 4$); a per-feature rescaling changes no axis-aligned tree, so under uniform truncation the two representations yield near-identical forests (measured max prediction difference 2-5e-3 at $m \leq 3$, against ~1e-1 at $m \geq 4$). Uniform truncation also removes the population's length *variance*, and Figure 4's distortion is a *between*-lengths quantity. E2 therefore separates two readings its own specification had conflated. Shortness per se costs the DCT nothing: refuted. What the data leave standing is the variance form: the drift coefficient's scale varies with length, so *mixed-length populations* are miscalibrated across dialogues. That form predicts exactly E2's null and exactly E1's localization. (One stray descriptive cell, MWOZ at $m = 8$ at -2.4 points nominal, is wrong-signed for either reading and is reported as such.)

E3, interventional, heterogeneous truncation (USS). The corrected causal instrument, designed after E2 was read: per-dialogue truncation depths drawn once from $\{2..8\}$ (seeded, shared across bases and head seeds), manufacturing length variance at fixed content and matched mean depth. The endpoint is inconclusive: $\Delta_{\text{het}} = +0.6$ points, positive but inside the design's stated resolution (the test resolves effects of roughly >= 1.5 points), and usable in neither direction. A lever-arm limitation is stated with it: truncation to 2-8 turns removes most of USS's label signal (accuracies fall to 0.55-0.67 from 0.76-0.78 untruncated), so the coefficients whose scale the manipulation miscalibrates have little signal left to corrupt. The MWOZ secondary replicate is null seed-ensembled (+0.0 points) but contains one wrong-signed nominal cell at single seed (seed 42: -4.9 points nominal), reported, like E2's stray cell, rather than explained away.

Figure 5 draws the two results that are shapes rather than single numbers.

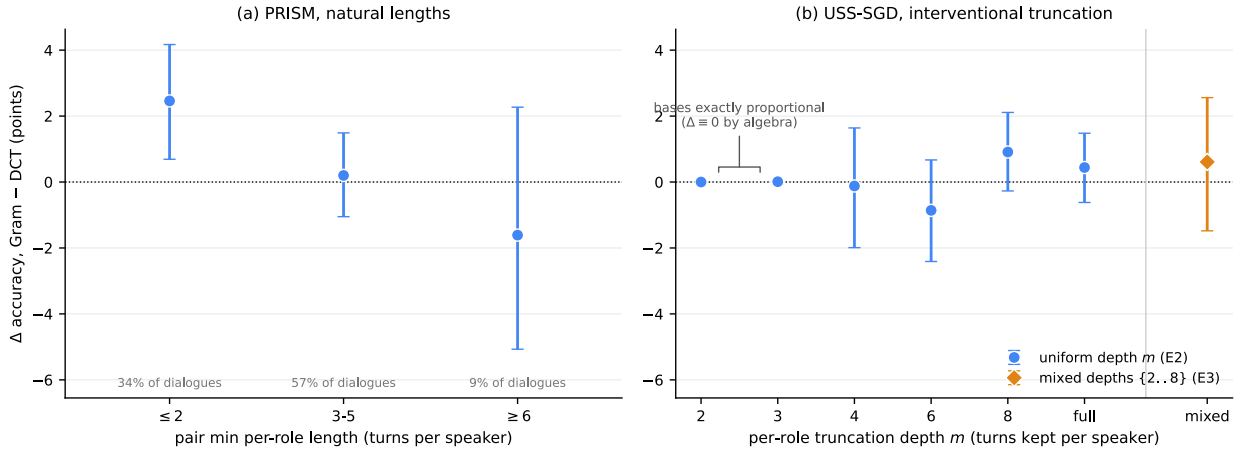

Figure 5: the Section 5 tests, drawn (seed-ensembled Δ = Gram − DCT in accuracy points, 95% bootstrap CIs). (a) PRISM at natural lengths: Δ by pair min per-role length, the exploratory tertiles; the primary endpoint that passes its rule is the $\leq$ 2-vs-$\geq$ 3 contrast of Table 2. The whole gap sits in the shortest stratum, which holds a third of the corpus's dialogues; the ≥ 6 stratum is 0.8% of pairs, hence its width. (b) USS-SGD under uniform truncation (E2, blue): null at every depth m, and at $m \leq 3$ the intervals collapse to hairlines. The two bases are exactly proportional feature-by-feature

there, so $\Delta \equiv 0$ by algebra, visible as vanishing error bars. The heterogeneous arm (E3, orange diamond) restores mixed lengths at matched mean depth and is inconclusive at this resolution.

What may now be claimed. Under the stated decision rules the exception passes as a localized property of very short per-role sequences, at a margin of +2.3 points. Its *mechanism*, the length-dependent scale of the DCT's drift coefficient, is supported by the localization it predicted and by exact algebra at the $T \leq 3$ boundary; its causal demonstration was attempted twice and remains unresolved. It is a supported hypothesis, not a demonstrated law. The decision table covers this outcome pattern (E1's rule passed, E2's did not) and prescribes its own suspicion, which stands: short PRISM chats may differ in content or label construct, and E1, an observational stratification on the corpus that generated the hypothesis, cannot exclude such a confound. For practice the asymmetry is one-sided. Nothing here weakens the tie at conversational lengths, and populations dominated by one-to-two-turn-per-role exchanges (short first-person chat) have a measured, mechanism-consistent reason to prefer the polynomial basis.

---

6. The Boundary of the Class: the DST Pays Only Where Shape Matters

What does leaving the class entirely cost? The DST is the natural probe of what the equivalence class requires: equally orthogonal, equally smooth, but with no constant basis function; its implicit odd extension pins the signal to zero at the boundaries.

Does the DST lose the mean? It has no constant basis function, so one might expect the truncated DST to collapse toward shape-only performance, the mean (the dominant signal channel) leaking across sine harmonics with ∼1/j decay. It does not: the first four DST harmonics capture 0.90-0.92 of the constant's energy at conversation lengths (the first half-sine alone carries $8/\pi^2 \approx 0.81$), so the mean survives truncation.

Table 3. Leaving the class: the DST costs accuracy only where shape matters.

basis, layout	USS-SGD	USS-MWOZ	PRISM
Gram, interleaved (reference)	0.719	**0.774**	**0.614**
DST, interleaved (sine harmonics 1-4, no mean)	0.726	0.740	0.613
DST + restored mean (Gram G_0 + DST harmonics 1-3)	**0.727**	0.758	0.611
shape-only (Gram degrees 1-3, mean removed)	0.706	0.727	0.524

Note: Seed-ensembled pairwise-preference accuracy at rating-gap >= 1.0, at matched dimensionality. The DST has no constant basis function, so its coefficients carry no mean on their own. The two reference rows bracket the range: the full Gram representation at the top, and a shape-only arm with the mean removed at the bottom. Bold marks the best score in each corpus column, a maximum marker at the table's precision rather than a significance statement.

The DST does not collapse: it sits at the Gram level on SGD and PRISM and far above shape-only on PRISM (+8.9 points, decided). It fails in exactly one place: MWOZ, where it loses 3.5 points, decided. MWOZ is the one corpus where trajectory shape carries *marginal* label signal beyond the mean: the marginal shape contribution measured on the same corpora (the interleaved arm minus mean pooling, five-seed) is -2.4 / +2.7 / -0.4 points on SGD/MWOZ/PRISM. The mechanism, consistent across all three corpora:

cost of leaving the DC-plus-smooth class $\sim \max(0, \text{marginal shape contribution})$.

Restoring the mean explicitly (the *DST + restored mean* row: Gram G_0 with DST harmonics 1-3) recovers only part of the MWOZ penalty, so the DC deficit is not the whole story: the sine shape directions themselves fit drifting trajectories poorly where shape matters. Within the class, basis choice is free; leaving the class costs precisely where trajectory shape is informative, and nothing where it is not.

---

7. Practical Guidance

1. Within the DC-plus-smooth class, let interpretability choose, because accuracy cannot: the bases tie at the instrument's resolution, so the readable basis costs nothing. For drifting, aperiodic sequences that is the polynomial one: G_0 = mean exactly, G_1 = the drift, per-coefficient semantics. For genuinely periodic signals the cosine basis's compaction argument returns. Where accuracy finally does discriminate, on populations dominated by very short per-role sequences (one to two turns per role, short first-person chat), it picks the same basis, by a measured margin (Section 5). Nothing is sacrificed in either regime.
2. Know your marginal shape contribution before leaving the class. One cheap ablation (full block vs mean pooling) prices what any shape-mangling basis or aggressive truncation can cost on your corpus.
3. Structural design beats basis design. The largest representation effect measured here, splitting the projection by speaker (Section 2.3), transfers across bases; invest design effort in the representation's structure before its basis.
4. Costs are a non-issue within the class, in either direction. Efficiency is sometimes offered as the reason to default to the DCT; at truncation it cannot be: truncated Gram and DCT are the same matmul shape, time, and space ($O(T \cdot d \cdot K)$ and $O(K \cdot d)$; the only per-call gap is basis construction, cacheable per length), and the DCT's celebrated $O(T \log T)$ advantage applies only to full spectra that sequence representations never compute. Both are dominated end-to-end by the embedding step, which outweighs either projection by a measured 780–$22{,}000\times$ across recorded hosts (medians, IQRs, and conditions are in the gram_timing artifact and its run history). Every representation compared in this paper is, end to end, the same cost.

---

8. Limitations

- One task family (conversation-level satisfaction as pairwise preference) and one embedding family for the basis experiments (MiniLM).
- Truncation at degree 3: the interleaved degree sweep (Figure 2) shows added degrees never improve on this task and corpus while residual energy keeps falling; other layouts, tasks, and corpora may want more.
- Instrument resolution of $\sim\pm 2.5$ points: the tie is an estimation statement at that resolution, not a proof of exact equality.
- The cross-length localization is confirmed at the instrument's stated resolution (Section 5); the mechanism's causal demonstration is not. One interventional design was algebraically unable to detect the mechanism's variance form, and the corrected design was under-powered on this corpus family; the mechanism remains a supported hypothesis.
- The confounding alternative the decision table names for this outcome pattern, content or label-construct differences on very short PRISM chats, remains unexcluded.

---

9. Conclusion

Gram projection gives sequence representation the polynomial basis done right: exact mean recovery, anytime truncation, readable coefficients. Measured head-to-head, nothing is paid for those properties. The DCT equivalent ties it at every conversational length, structural gains transfer between the bases, and the one located exception (populations mixing very short per-role sequences) favors the readable basis by a measured margin, with the same-corpus caveat Section 5 states. The basis outside the class breaks the pattern by a mechanism you can price in advance. Choose the basis you can read, and spend your design effort on what you project.

Appendix A. Reproduction Kit

The reproduction kit is public at `github.com/suisuss/gp-paper`: two tracked, portable `lib.pipeline` specs plus the standalone generator scripts it enumerates, regenerating every artifact from public data (USS, PRISM) with one command (`./run.sh`). A fast `./run.sh check` validates the paper against its shipped artifacts without reproducing them: every table cell matches its source artifact, every figure and cited result names the script that produced it, every artifact is provenance-clean (git commit, script sha256, and seeds recorded, all built from a clean checkout), and the committed PDF rebuilds from source with matching content. These gates also run in continuous integration on every push and pull request, so the paper cannot silently drift from its artifacts. The kit's claim-to-script-to-artifact map covers every generated table (the identities, basis-tie, localization, and DST-cost tables, each cell checked against its source artifact) and all five figures. The Section 5 tests follow an explicit specification of endpoints, gates, and decision rules, and every arm's outcome is reported under it.

Appendix B. Numerical Verification of the Identities

The Section 2.1 identities, verified on the corpora this paper evaluates; cells give the max absolute error over every dialogue in all three corpora.

check	float64	float32
G_0 vs the arithmetic mean	5.6e-17	2.9e-08
degree-independence ($K = 1$ vs $K = 3$ fits, shared degrees)	0 (exact)	0 (exact)
midline recovery $(\text{fit}(U) + \text{fit}(M))/2$	5.6e-17	7.8e-09
offset recovery $\text{fit}(U) - \text{fit}(M)$	5.6e-17	9.3e-09

(Recovery checks run on strictly-alternating dialogues with equal per-role turn counts: 1,000 on USS-SGD, 7,738 on PRISM; USS-MWOZ contains none. float32 errors sit at that precision's representational floor; the identities cost nothing.)

References

- Ahmed, N., Natarajan, T., and Rao, K. R. (1974). Discrete Cosine Transform. *IEEE Transactions on Computers*, C-23(1):90-93.

- Almarwani, N., Aldarmaki, H., and Diab, M. (2019). Efficient Sentence Embedding using Discrete Cosine Transform. *EMNLP-IJCNLP 2019*.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5-32.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ — A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *EMNLP 2018*.
- Hosking, J. R. M. (1990). L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society, Series B*, 52(1):105-124.
- Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. (2024). The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *NeurIPS 2024 Datasets and Benchmarks* (arXiv:2404.16019).
- Ramsay, J. O. and Silverman, B. W. (2005). Functional Data Analysis (2nd ed.). *Springer*.
- Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *AAAI 2020*.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019*.
- Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627-1639.
- Sun, W., Zhang, S., Balog, K., Ren, Z., Ren, P., Chen, Z., and de Rijke, M. (2021). Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems. *SIGIR 2021* (arXiv:2105.03748). (USS.)
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *NeurIPS 2020*.