SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

DOES SCRIPT CONDITIONING HELP DYSLEXIA DETECTION? A CAPACITY-MATCHED ABLATION ON URDU–ENGLISH HANDWRITING

Ayesha Ashfaq[1], Abdullah Butt[*2]

[1]*Lahore, Pakistan*
[*2]*United Arab Emirates*

[1]ayeshaashfaq925@gmail.com, [*2]abutt2210@gmail.com

[1]ORCID: 0009-0002-1180-8737
[*2]ORCID: 0009-0000-7724-890X

DOI:https://doi.org/10.5281/zenodo.22689149

Keywords
dyslexia screening; handwriting classification; Urdu; script conditioning; dataset integrity; null result; pre-registration

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Corresponding Author: *
Abdullah Butt

Abstract

Handwriting images are an unusually accessible signal for early dyslexia screening, and the recent release of a bilingual Urdu–English children's handwriting corpus makes it possible to ask whether a model should be told which script it is looking at. We test that question and answer it in the negative, and we report first what we had to establish before the question could be asked at all. A content-level audit of the released corpus finds that its 852 files hold only 638 distinct images: one class folder contains 194 redundant copies, 20 images are byte-identical across both diagnostic folders and therefore carry contradictory labels, and the corpus's stated one-to-one class balance holds only at the file level. Deduplicated and with the contradictions removed, the corpus is 406 dyslexic against 212 non-dyslexic images, and the majority-class baseline moves from 0.500 to 0.657. On that corrected corpus we run a pre-registered, capacity-matched ablation: a script-conditioned channel-attention module against an identical-size control whose gates receive a learned constant instead of the script, under one fixed split, 30 shared seeds, two co-primary endpoints and a freeze tag applied before any test evaluation. Neither endpoint is rejected. The pooled accuracy difference is +0.0075 with a 95% confidence interval of [−0.0174, +0.0325], and the Urdu-subset difference is +0.0008 with an interval of [−0.0216, +0.0232]. We report the achieved power of this design honestly: it is 0.121 for the pre-specified effect of 1.5 accuracy points, so the correct reading is that the design could not detect an effect of that size, not that no effect exists. A counterfactual probe of the attention gates supplies a mechanism for the null rather than leaving it unexplained: holding each image fixed and varying only the script identifier moves the gates about 3% as much as changing the image does. We claim no novelty for script-aware computation as a mechanism. The contribution is the audit, the controlled test, and the mechanism, in that order.

1. INTRODUCTION

Children with dyslexia frequently produce handwriting with irregular spacing, letter reversals and inconsistent character sizing, and the transcription difficulties behind those patterns persist into adulthood [1]. Handwriting can be

collected with paper, a pen and a phone camera. That combination has made handwriting images an attractive signal for early screening support in settings where specialist assessment is out of reach, and a body of work now applies deep learning to them [2–4]. Nearly all of it is monolingual. For Urdu, a low-resource language written in a cursive, dot-bearing Nastaliq style, a bilingual children's handwriting corpus for dyslexia detection has appeared only recently [5].

That corpus invites an obvious design question. Urdu and English character images have sharply different visual statistics: Urdu characters are cursive, carry identity-bearing dot configurations and vary with ligature context, while English characters are discrete Latin glyphs. A script-agnostic backbone must serve both distributions with shared capacity. Would telling the model which script it is looking at help? This paper answers no, and reports two things a reader should know from the first page.

1.1 First, the corpus is not what its file counts suggest

Before any model was trained we computed SHA-256 digests over the release. The 852 files hold **638 distinct images**. One class folder is internally clean; the other contains 426 files but only 232 distinct images. Twenty images are byte-identical across both class folders and therefore carry contradictory labels: the same pixels are filed as dyslexic and as non-dyslexic. The stated one-to-one class balance is an artefact of duplication. Deduplicated, and with the contradictory images removed, the corpus is **406 dyslexic against 212 non-dyslexic**, so the majority-class accuracy a model must beat is 0.657, not 0.500. Any image-level split of the released files puts pixel-identical images on both sides of the partition. We report the audit in full in Section 3, and we regard it as this paper's most portable contribution: the procedure and the reporting convention transfer to any released corpus, whatever happens to this one.

1.2 Second, script conditioning does not measurably help, and we can say why

On the corrected corpus we compare a script-conditioned channel-attention module (arm A2) against a control of identical parameter count whose gates receive a learned constant vector in place of the script embedding (arm A1), under one fixed split, an identical tuning budget, 30 shared random seeds, and two co-primary endpoints frozen in the repository before any test-set evaluation. Neither endpoint is rejected. The pooled accuracy difference is +0.0075, 95% CI [−0.0174, +0.0325]; the Urdu-subset difference, where the script-statistics argument predicts any benefit should concentrate, is +0.0008, 95% CI [−0.0216, +0.0232].

We state the limit of that result in the same breath as the result. The design's achieved power for its own pre-specified minimal effect of 1.5 accuracy points is **0.121**. The smallest effect it could detect at 80% power is 4.33 accuracy points. *This design could not detect an effect of the pre-specified size.* It is not evidence that no effect exists. The confidence intervals above, not the *p*-values, carry what was learned, and they exclude an advantage larger than about 3.3 accuracy points.

A null at this power would be thin on its own. What makes it informative is a direct measurement of the mechanism. We probe the attention gates counterfactually: for each test image we hold the pixels fixed and vary only the script identifier through all three values, reading the gate vector each time. Changing the script moves the gates with a standard deviation of 2.82×10^{-4}, against 8.61×10^{-3} for changing the image: a ratio of **0.0327**. The conditioning signal barely perturbs the attention it was introduced to steer. That is a mechanistic account of the null rather than a restatement of it.

1.3 Contributions

1. **A content-level integrity audit of a released bilingual dyslexia handwriting corpus** (Section 3): its duplicate structure, its cross-class label contradictions, the leakage consequence for any image-level split, and the resulting correction to the stated class balance. The audit is exact-hash and therefore a lower bound on redundancy.

2. **A pre-registered, capacity-matched, multi-seed ablation of script-conditioned attention** against a matched-capacity control on an identical fixed split, with two co-primary endpoints, Holm correction, a freeze gate before any test evaluation, and the null reported at full prominence with its achieved power stated (Sections 5 to 7).
3. **A mechanistic explanation of the null** via a script counterfactual on the attention gates, validated by a null check on the control arm (Section 8).
4. **An internally consistent re-run of the source study's backbones** under one fixed split, including two arms that collapsed to constant majority-class predictors and are reported as such rather than dropped (Section 9).
5. **A released script-annotation layer** for the corpus, with its agreement statistic described accurately as human-versus-AI and its adjudication as human-authoritative (Section 4).

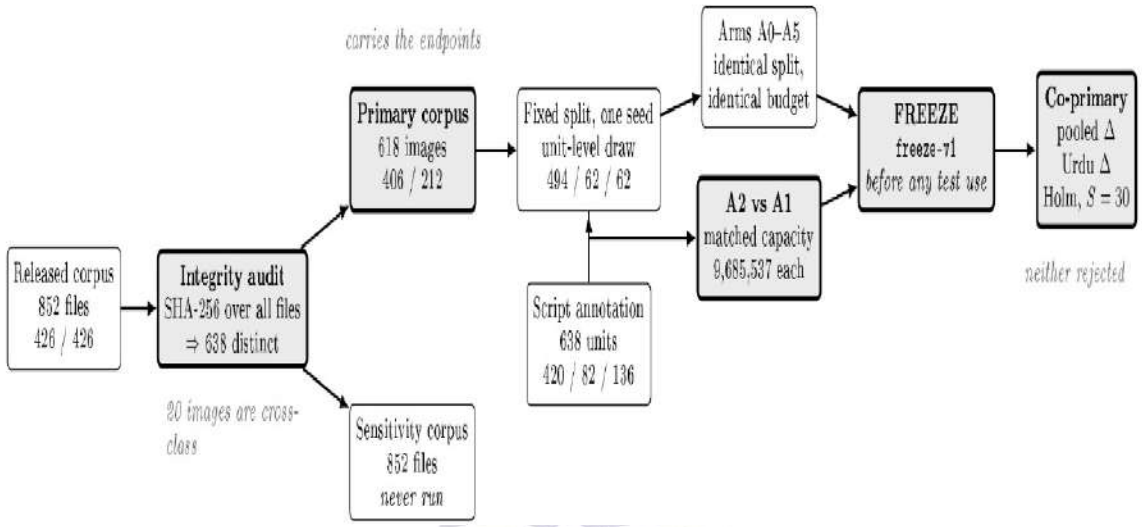

Figure 1. The study in one figure. Shaded boxes are the load-bearing steps. The audit is upstream of everything: the corpus the experiment runs on does not exist until the released files have been checksummed and the duplicates and label contradictions removed. The freeze gate precedes any use of the test partition, and only the two co-primary endpoints carry inferential claims.

1.4 What we do not claim

We claim no novelty for the mechanism of conditioning computation on script; it is established in multi-script recognition, where script identity determines the label space and conditioning is structurally necessary. Our setting differs in that the label space is shared across scripts, so whether conditioning helps is an empirical question rather than a structural requirement, and that question is what we answer. Any statement in this paper about what has or has not been evaluated before rests on literature searches conducted in September 2026, whose queries and dates are recorded in the supplement; we make no claim about work appearing after that date. Absolute accuracies here are within-dataset and are not generalisation estimates; we make no screening or deployment claim.

2. Related Work

Every source cited below was retrieved and read. Where we characterise a paper, the characterisation comes from its text, and where our access was limited to a registered abstract we say so at that point.

2.1 Handwriting-based screening for dyslexia and dysgraphia

Automated analysis of children's handwriting for learning-disorder screening spans dysgraphia and dyslexia and a range of capture modalities, from tablet and sensor-pen kinematics to offline images; [2] survey the area. The online modalities are the

higher-signal ones. [6] assess developmental dysgraphia from handwriting recorded on a display tablet, and [7] use an instrumented pen. Both capture temporal and pressure information that an offline photograph discards. The present study works in the offline-image setting, which is the more accessible and the harder one, and our modest absolute accuracies should be read against that.

For offline images in Latin script, [3] combine transfer learning with attention-era architectures and Grad-CAM visualisation, reporting a test precision of 99.65%. Figures of that magnitude are common in this literature and warrant care about what is being classified: corpora labelled at the letter level for orientation (normal, reversal, corrected) encode what a letter looks like, not who wrote it, which is a different construct from a child-level dyslexia label.

2.2 Attention, and the split that is possible

Attention has entered dyslexia detection, but not as script conditioning. [4] combine a convolutional feature extractor with positional encoding, a bidirectional LSTM and attention over the temporal sequence of a Chinese dictation task, training on handwriting images of 100,000 Chinese characters from 1,064 children in Grades 2 to 6 and reaching 83.2% accuracy from handwriting features alone. One methodological detail of that study matters here more than its accuracy. Their training, validation and test sets were divided "in a stratified grouping of both grade and dyslexic status, resulting in a sample size of 851:106:107" [4]; those counts sum to 1,064, the number of children. Their split is at the participant level. The release used here carries no writer identifiers at all, so no such split is possible, which is the central limitation of the present work (Section 11).

2.3 The bilingual corpus this study uses

The closest prior work is [5], who introduced the bilingual English–Urdu children's handwriting corpus analysed here. Their published abstract describes a framework for detecting dyslexia from handwritten samples of children aged 6 to 10, using a balanced English–Urdu dataset of 852 samples, and reports that among the deep architectures they evaluated, MobileNetV1 with the RMSProp optimiser performed best, with an F_1 of 0.7823.

We must be explicit about the limits of that description. The publisher's full text was not retrievable to us: both the article and PDF endpoints returned HTTP 403. The sentence above is drawn from the registered Crossref and OpenAlex records for the article's DOI, which we retrieved, and not from the paper's body. We therefore do not characterise their methodology, their split, their preprocessing or their per-architecture results anywhere in this paper, and where our own results would ordinarily be set against theirs we leave the comparison unmade rather than reconstructed (Section 9).

What we can say is architectural and follows from the corpus itself: every model in that benchmark is script-agnostic, in that Urdu characters, English characters and digits pass through one shared network with no script information. We adopt their dataset, audit it (Section 3), re-run a comparable set of backbones under our own fixed split for internal consistency, and add exactly one design factor, script conditioning, under matched capacity.

2.4 Urdu and Arabic-script resources

Urdu remains under-resourced relative to Latin-script settings, and dataset construction is active: [8] introduce a multi-task Urdu scene-text corpus with detection, recognition and visual question answering annotations. The task and modality are remote from ours, and we cite it only to situate Urdu resource availability. [9] assess Arabic-speaking children with dyslexia using a standardised language battery. That work is relevant here not as a computational baseline but as a point of contrast on construct validity: it illustrates what a standardised identification procedure looks like, against which the institutional labels of the present corpus should be read (Section 11).

2.5 Script-aware computation

Script-aware processing is not a new mechanism. Multi-script handwriting recognition has combined script identification with recognition, either as cascaded stages or jointly: [10] present multilingual text recognition networks performing simultaneous script identification and handwriting recognition in one model. We claim no novelty for the idea of conditioning computation on script.

The distinction that matters for the present study is structural. In recognition, script identity determines the label space, so script-awareness is a requirement. In binary dyslexia classification the label space is shared across scripts, so whether script conditioning helps is an open empirical question. That is the question this paper tests, and the answer it obtains is negative.

3. The Release and Its Integrity Audit

We use the bilingual children's handwriting corpus of [5], obtained through the access route published in that paper on 2026-08-31, and we do not redistribute the images. The folder listing as accessed and per-file SHA-256 digests are released with our code.

The registered record for that article describes a dataset of 852 handwritten samples from children aged 6 to 10, balanced between English and Urdu [5]. We are able to confirm the file counts and the two-class folder structure directly from the release. Beyond that we describe the corpus only as we received it, because the published full text was not accessible to us (Section 2), and we prefer an incomplete description to a reconstructed one. Readers who need the collection protocol, the recruitment procedure or the ethics record should consult the source article, which is open access under CC BY 4.0.

This limitation is real and we do not minimise it: a secondary analysis that cannot read its source study's methods section is working with less context than it should have. It does not affect the internal validity of what follows, because every comparison in this paper is between arms trained and evaluated on the identical partition of the identical files, but it does bound what we can say about how those files came to exist.

3.1 What the guide-mandated check passed, and what it missed

The release contains 852 JPEG files, 426 in each class folder. A verification pass on those counts succeeds, and had we stopped there the study would have proceeded on a corpus we did not understand. We additionally computed a SHA-256 digest for every file. The counts do not survive it.

Table 1 gives the per-folder position and Table 2 gives the audit in full. The 852 files hold 638 distinct images. The dyslexic folder is internally clean: 426 files, 426 distinct images. The non-dyslexic folder contains 426 files and 232 distinct images, so 194 of its files are redundant copies. Across the release there are 147 duplicate groups, and 20 of them span both class folders: the same bytes appear under both diagnoses.

Class folder	Released files	Distinct images	Redundant copies
Yes/ (dyslexic)	426	426	0
No/ (non-dyslexic)	426	232	194
Total	**852**	**638**	**214**

Table 1. Content-level audit of the released corpus. File counts match the release exactly; content counts do not. The stated 1:1 class balance survives only at the file level and is produced by duplication inside No/. The Total row is a release-level count and not a column sum: distinct images total 638 rather than 658 because the 20 cross-class images are counted once, and redundant copies total 214 rather than 194 because those 20 cross-class occurrences are redundancy at the release level. Generated from results/duplicate_audit.json and data/corpus_v1_summary.json.

Audit quantity	Value
Duplicate groups (SHA-256)	147
of size 2 / 3 / 4	82 / 63 / 2
Groups spanning both class folders	**20**
Groups within one class folder	127 (all in No/)
Primary corpus (deduplicated, contradictions removed)	**618**
dyslexic / non-dyslexic	406 / 212
class ratio	1.9151:1
Sensitivity corpus (released files as-is)	**852**

Table 2. The integrity audit in full. Group sizes reconcile with the redundancy total: 82 + (63 × 2) + (2 × 3) = 214 redundant copies, and 20 + 127 = 147 duplicate groups. Generated from results/duplicate_audit.json.

3.2 Three consequences

Leakage. With an image-level split and 147 duplicate groups, a group straddling the train and test partitions places a pixel-identical image on both sides. Absolute accuracy is then inflated by memorisation rather than generalisation. This applies to any image-level evaluation of the released files, ours and any other.

Label contradiction. Twenty images are simultaneously labelled dyslexic and non-dyslexic. No classifier can be correct on both copies. They impose an irreducible error floor on any partition they occupy and corrupt any partition they straddle.

The stated class balance is an artefact. The 426/426 balance holds at the file level only, and is produced by duplication inside the non-dyslexic folder. Deduplicated and with the 20 contradictory images removed, the corpus is 406 dyslexic against 212 non-dyslexic, a ratio of 1.92 to 1. A model that predicts the majority class for every image scores 0.657, not 0.500. Every absolute accuracy in this paper must be read against that figure, and against 0.6613, the corresponding rate on our particular test partition.

3.3 Scope and honesty of the audit

The audit is *exact-hash*. It detects byte-identical files and nothing else. We also ran a perceptual near-duplicate scan; on low-entropy single-character images it chained unrelated images into incoherent clusters and merged images of differing dimensions, so we discarded it and report only the exact-hash result, which is definitive as far as it goes. It follows that 638 is an *upper bound* on the number of distinct images and 214 a *lower bound* on redundancy: a re-encoded or rescaled copy of the same handwriting sample hashes differently and would survive our deduplication undetected. We state this rather than presenting the audit as complete.

We do not characterise this as an error by the source authors, whose file counts are exactly as described. We characterise it as a property of the release that any user must check for, and that we found only because we checksummed rather than counted. The generalisable claim is that file counts are not content counts, and that a published corpus should report both.

4. Corpus Definitions, Script Annotation, and the Fixed Split

4.1 Two corpora

The audit forces a choice, which we resolved before any modelling. The **primary corpus** is the deduplicated release with the contradictory images removed: 618 images, 406 dyslexic and 212 non-

dyslexic, one representative per distinct digest chosen deterministically as the lexicographically smallest record key in each group. This corpus carries the co-primary endpoints. The **sensitivity corpus** is the 852 released files as-is, retained so that comparability with the source study is not lost.

We must report that the sensitivity corpus was defined, split, and subjected to its integrity assertions, but was **never run**. Every one of the 472 logged runs uses the primary corpus. The sensitivity analysis appears in the frozen analysis plan and was not executed; we report it as not executed rather than omitting it.

4.2 Script annotation

The release provides the diagnosis label only; script is not recorded. Every image was therefore annotated with a script tag in urdu, english, digit. The annotation unit is the *distinct image* rather than the file, so 638 units were annotated and every duplicate file inherits its unit's tag; this makes inconsistent tags across byte-identical files impossible. Annotators worked from the image alone, on contact sheets bearing only an opaque unit identifier, in an order shuffled with a recorded seed so that class-folder structure could not leak through presentation order.

Two properties of this annotation must be stated plainly, because both depart from what a reader would assume.

The agreement statistic is human-versus-AI. Annotator A was an AI coding assistant working from images alone; annotator B was a human author. Cohen's κ on the script tag is **0.8489** with 92.13% raw agreement over the 635 units both annotators judged. This is *not* an inter-human reliability figure and must not be read as one.

Adjudication was human-authoritative, not joint. Every one of the 50 disagreements was resolved in favour of the human annotator by a policy fixed before the stage was run. The final script column therefore reproduces the human annotator's judgement wherever one was given, and the released file is human-adjudicated in that specific sense rather than through case-by-case joint discussion.

The adjudicated composition is 420 Urdu (65.8%), 82 English (12.9%) and 136 digits (21.3%). Among the digits, **135 are Western numerals and none are Eastern Arabic-Indic forms** (one is unreadable). This is consequential rather than incidental: the corpus contains no Urdu-form numerals, so the script-statistics argument that motivates this study makes no prediction at all for the digit category, and digit results should be read as a third visual class rather than as a third script. It also means the Arabic handwritten-digit recognition literature, which works on Eastern Arabic-Indic forms and corpora such as MADBase [11], does not bear on the digit results reported here.

One further provenance detail belongs in the open. The released annotation file has *mixed provenance*: its script column is the human annotator's, while its ambiguity flag column is annotator A's in all 638 rows. The AI annotator flagged 117 units (18.3%) as ambiguous between scripts; the human annotator flagged 4. We report both figures rather than the aggregate, because an aggregate over "at least one annotator" would attribute an AI-generated rate to a human-adjudicated file.

4.3 The fixed split

One split was generated, with a recorded seed, and reused unchanged by every arm. It is drawn at the level of the annotation unit rather than the file, so every byte-identical copy of an image lands in a single partition and no duplicate group straddles a boundary; this is asserted in the split generator and re-checked in the test suite. Partitions are stratified jointly on class and script, in an 80/10/10 ratio, giving 494 training, 62 validation and 62 test images. The test partition holds 41 dyslexic and 21 non-dyslexic images, and by script 43 Urdu, 8 English and 11 digits.

Those subset sizes are small enough to govern how the results may be read. On the pooled test partition one image is worth 1.61 accuracy points. On the Urdu subset one image is worth 2.33 points, on the digit subset 9.09, and on the English subset 12.5. We return to this in Section 11.

5. Method

5.1 Design principle

Every arm differs from its comparator by exactly one factor, and the primary comparison holds parameter count fixed while varying only the *information* available to the model. All arms use a frozen ImageNet-pretrained MobileNetV1 backbone at 224 × 224 × 3 with the standard MobileNet preprocessing, with small trainable modules on top. No augmentation, no synthetic data and no external data are used.

5.2 The script-conditioned attention module

The description below is taken from the code that ran, not from a specification. Let $F \in \mathbb{R}^{H\times W\times C}$ be the feature map at an insertion depth and let d be the script-embedding width. The module computes:

$$z = \frac{1}{HW}\sum_{h,w} F_{h,w,:} \in \mathbb{R}^C, \tag{1}$$

$$e = \begin{cases} \text{Embed}(s) \in \mathbb{R}^d & \text{conditioned (A2)} \\ c \in \mathbb{R}^d & \text{unconditioned (A1)} \end{cases} \tag{2}$$

$$\alpha = \sigma\big(W_2\,\text{ReLU}(W_1[z;e])\big) \in (0,1)^C, \tag{3}$$

$$F' = F + F \odot \alpha = F \odot (1+\alpha), \tag{4}$$

where W_1 maps to max(C/r, 4) units, σ is the sigmoid, and c is a trainable constant vector initialised to zero. The gate α is broadcast identically across every spatial position.

There is no spatial term anywhere in the mechanism. This is a *channel* gate in the squeeze-and-excitation style of [12], with the script embedding concatenated to the pooled descriptor before the bottleneck so that the recalibration is conditioned rather than purely self-derived. It matters for how the mechanism can be visualised (Section 8) and for what it can plausibly do: the module can reweight feature channels per script, but it cannot attend to a region of a character.

Matched capacity, stated precisely. Both variants instantiate both the embedding table and the constant vector, so total and trainable parameter counts are identical by construction: at the insertion depth used for the co-primary comparison both arms have exactly 9,685,537 parameters, and equality was asserted at all 16 configurations of depth, embedding width and bottleneck ratio. We state one qualification that a parameter count cannot see. In the conditioned arm the constant vector receives no gradient; in the control the embedding table receives none. The *gradient-receiving* counts therefore differ by $2d$ = 32 parameters, which is 5×10^{-4} percent of trainable capacity. Capacity is matched to the parameter; it is not matched to the last degree of freedom, and we prefer to say so.

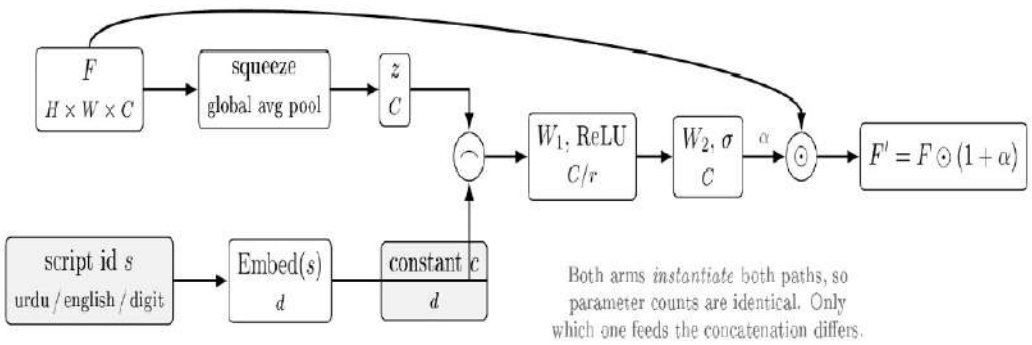

Figure 2. The script-conditioned attention module, drawn from src/models.py. The gate α is a channel vector broadcast across all spatial positions; the module has no spatial term. The matched-capacity control (A1) is the same graph with the constant path selected in place of the embedding path, which is why the two arms differ in information and not in size.

5.3 Experimental arms

Table 3 specifies every arm. The confirmatory pair is **A2 against A1**. All other arms are context.

Arm A3 exists to separate the value of the script *information* from the value of the attention *mechanism* that consumes it. It carries no attention module at all and simply concatenates a one-hot script vector onto the penultimate features, in the spirit of the feature-wise conditioning layers of [13], though without their affine modulation.

Three points about the context arms are easy to misread and are stated here rather than in a footnote.

- **A5 and A4 are routed by the oracle script tag, not by a classifier.** A5 selects among three script-specific heads and A4 among two script-specific expert models, and in both cases the routing variable at inference is the adjudicated annotation, not a prediction. Neither is a deployable pipeline, and neither should be read as one.
- **A2′ is the only arm in the study driven by a predicted script**, at both training and inference, from a classifier trained on the same training partition.
- **A4's digits are routed to the English expert.** Since the corpus contains no Eastern Arabic-Indic numerals, all digits are Western-form, and the routing rule sends every non-Urdu image to the English expert. That expert never saw a digit in training. A4's digit performance is a consequence of this design and is reported as such.

Arm	What it changes	Role	Script id	Params
A0	Source-study head on a frozen backbone; script input ignored	context	oracle (unused)	9,651,649
A1	SCA gates see a learned *constant*: matched-capacity control	**confirmatory**	n/a	9,685,537
A2	SCA gates see the script embedding	**confirmatory**	oracle	9,685,537
A2′	A2 with script ids from a trained classifier	context	predicted	9,685,537
A3	One-hot script concatenated to penultimate features; no SCA	context	oracle	9,651,652
A4	Separate Urdu and English experts, routed at inference	context	oracle	19,303,298
A5	Three script-specific heads over shared frozen features	context	oracle	22,497,219

Table 3. Experimental arms. Only the A2–A1 pair is confirmatory. Parameter counts are read from the run log. A1 and A2 are identical in size by construction and differ only in whether the attention gates receive the script. The 'script id' column records what each arm was actually fed at inference; A2′ is the only arm in the study driven by a predicted script (see Section 5.3). Co-primary pair at the matched insertion depth mid: A1 rmsprop, lr 0.0003; A2 rmsprop, lr 0.0001; both 9,685,537 parameters (DEVIATIONS.md D-006). Generated from results/all_runs.csv and results/selected_configs.json.

6. Pre-Registered Protocol

6.1 Endpoints, correction and the freeze gate

Two co-primary endpoints were pre-specified: the pooled A2−A1 accuracy difference on the test partition, and the Urdu-subset A2−A1 accuracy difference. The Urdu endpoint exists because the script-statistics argument predicts that any benefit concentrates in Urdu, and a script-specific effect must not be averaged away into a flat pooled number. A Holm correction is applied across the two at a family-wise α of 0.05. Each endpoint is tested with a paired *t*-test over the seed-matched differences, with a Wilcoxon signed-rank test as a distribution-free check, Cohen's d_z, and a 95% confidence interval. Everything else is exploratory. The tuning grid, the endpoints, the selected configurations, the repeat count and the analysis plan were committed and tagged freeze-v1 before any test-partition evaluation, and the training code refuses a test-partition run unless the tag exists. Model selection used the validation partition only.

6.2 Matched insertion depth

Validation selection chose different insertion depths for the two arms, which would have left the compared models at 9,821,161 against 9,685,537 parameters, a 1.40% gap in the control's favour. Since the design's central claim is that the two arms differ only in information, we fixed the co-primary comparison at a single common depth, at which both arms have exactly 9,685,537 parameters, while each arm still selected its own optimiser and learning rate inside that depth so the search budget stayed equal. The chosen depth is the treatment arm's best and is tied-best for the control, and the validation difference is identical either way, so the decision fixed the integrity of the capacity claim without moving the number. It was taken before any test data was touched. The each-arm-own-best pairing is retained as an exploratory sensitivity analysis.

We note one residual asymmetry rather than leaving it for a reviewer: the two arms carry different learning rates within the common depth, as a consequence of tuning each independently. This is what keeps the search budget equal, but it means the confirmatory contrast is matched in capacity and not in every hyperparameter.

6.3 Power, stated before the result

A five-seed pilot on the co-primary pair, run on validation only, estimated the between-seed standard deviation of the paired accuracy difference at 0.0736. Achieving 80% power for the pre-specified minimal effect of 0.015 would have required 232 seeds. The pre-registered ceiling of 30 binds, so S = 30, at which the **achieved power for a 1.5-point effect is 0.121** and the smallest detectable effect at 80% power is **4.33 accuracy points**.

A second constraint is arithmetic and independent of S: the test partition holds 62 images, so one image is worth 1.61 accuracy points and the 1.5-point target lies *below the resolution of a single-seed measurement*. Averaging over seeds refines the mean; it does not reduce the seed-to-seed variance that drives power.

The interpretation rule was fixed in the repository before unblinding and is reproduced here unchanged: a null result is not evidence that script conditioning fails, only that no effect larger than roughly four accuracy points was detected; the confidence interval, not the *p*-value, carries the informative content.

6.4 A defect in the training loop, disclosed

We must report a defect that we discovered after all runs were complete.

The training routine passes its evaluation partition to an early-stopping callback that restores the best-scoring weights on exit. During tuning and the pilot the evaluation partition is the validation set and this is correct. In the final phase the evaluation partition is the test set. For six of the seven arms, therefore, **the stopping epoch** and the restored weights were selected by minimising loss on the same 62 test images that were then scored. The freeze gate did not catch this: it governs when the test partition may be touched, not how it is used once open.

Three consequences follow, and we prefer to draw them ourselves.

First, every absolute test figure for those six arms is optimistically biased by epoch selection. Table 7 is not a clean held-out measurement and is not presented as one.

Second, the confirmatory contrast is the most defensible quantity in the study, but this requires an argument rather than an assumption. The two arms are contaminated identically: the same 62 images, the same monitored quantity, the same patience, the same restoration rule, the same fixed split, the same seed list, and identical parameter counts. The contamination is a common-mode term in a paired difference. It does not cancel by construction, since epoch selection is per-run, but there is no mechanism by which it would systematically favour the conditioned arm; and per-run selection adds variance to the paired difference, which lowers power further. The bias is therefore conservative with respect to this paper's conclusion.

Third, one arm is exempt and is thereby made incomparable. A4 early-stops on same-script validation images and is the only arm produced under the protocol as intended. Its row in Table 7 was generated under a different rule from the other six and should not be compared with them. We report this rather than repairing it silently. The code is left unmodified so that it continues to match the runs that produced the results, and the deviation is logged in full in the supplement.

6.5 Reproducibility record

Table 6 records the frozen configuration under which every number in this paper was produced. Every table and figure is generated by one command from artifacts on disk; no number is hand-entered.

Item	Value
Fixed split seed	1337 (generated once, never regenerated)
Draw level	annotation unit (distinct image, SHA-256)
Partitions (primary)	494 / 62 / 62
Tuning seeds	101–103
Pilot seeds	201–205
Final seeds	301–330 (S = 30)
Freeze tag	freeze-v1 at commit e491aaf
Runs logged	472 (252 tuning, 10 pilot, 210 final)
Total compute	24.5 h
Environment	tf2.20.0 / keras3.15.1 / py3.13.7 / Windows / cpu
Hardware	Intel i7-6600U, 2 physical cores, 16 GB RAM
Duplicate audit	python src/audit_duplicates.py
Corpus definitions	python src/build_corpus.py
Fixed split	python src/make_split.py
Tuning sweep	python src/run_tuning.py --arms A1 A2
Pilot and power	python src/run_pilot.py
Final runs	python src/run_final.py --arms A1 A2 A0 A3 python src/run_final.py --arms A5 python src/run_final.py --arms A2p A4

Item	Value
Statistics (Table 4)	python src/stats.py
Gate probe (Figure 4)	python src/attention_probe.py
Backbone table (Table 8)	python src/a0_table.py
Figures and Table 7	python src/figures.py
Tables 2, 3, 4, 5, 6	python src/paper_tables.py

Table 6. Reproducibility record. Generated from results/all_runs.csv and data/splits/split_v1.json.

7. Results: Confirmatory

This section reports the two pre-registered co-primary endpoints and nothing else. No other comparison in this paper carries a significance claim.

Neither endpoint is rejected. Table 4 gives both. The pooled A2−A1 accuracy difference over 30 seed-matched pairs is +0.0075, with a 95% confidence interval of [−0.0174, +0.0325], d_z = +0.113, p_t = 0.5420 against a Holm threshold of 0.0250. The Urdu-subset difference is +0.0008, 95% CI [−0.0216, +0.0232], d_z = +0.013, p_t = 0.9441 against a threshold of 0.0500. The Wilcoxon signed-rank test agrees with the *t*-test on both endpoints. The conditioned arm reaches 0.7495 mean accuracy and the control 0.7419, both against a majority-class rate of 0.6613 on this partition.

Figure 3 shows both intervals with the pre-specified minimal effect drawn as a band on the same axis, so that what the interval excludes and what it does not are visible together.

Endpoint	A1	A2	Δ	95% CI	d_z	p_t	p_W	Holm α	Reject
Pooled	0.7419	0.7495	+0.0075	[−0.0174, +0.0325]	+0.113	0.5420	0.4168	0.0250	**no**
Urdu subset	0.7357	0.7364	+0.0008	[−0.0216, +0.0232]	+0.013	0.9441	0.9198	0.0500	**no**

Table 4. Co-primary endpoints. Seed-paired A2−A1 accuracy differences over S = 30 seeds on the fixed test partition, Holm-corrected across the two. Neither is rejected. The confidence interval, not the *p*-value, carries the informative content: at the achieved power of 0.121 this design could not detect an effect of the pre-specified size (δ_{min} = 0.015). Majority-class rate on this test partition: 41/62 = 0.6613. One test image is worth 0.0161 accuracy points. Smallest effect detectable at 80% power with S = 30: 0.0433. Achieving 80% power for δ_{min} would have required S = 232. Generated from results/stats_summary.json.

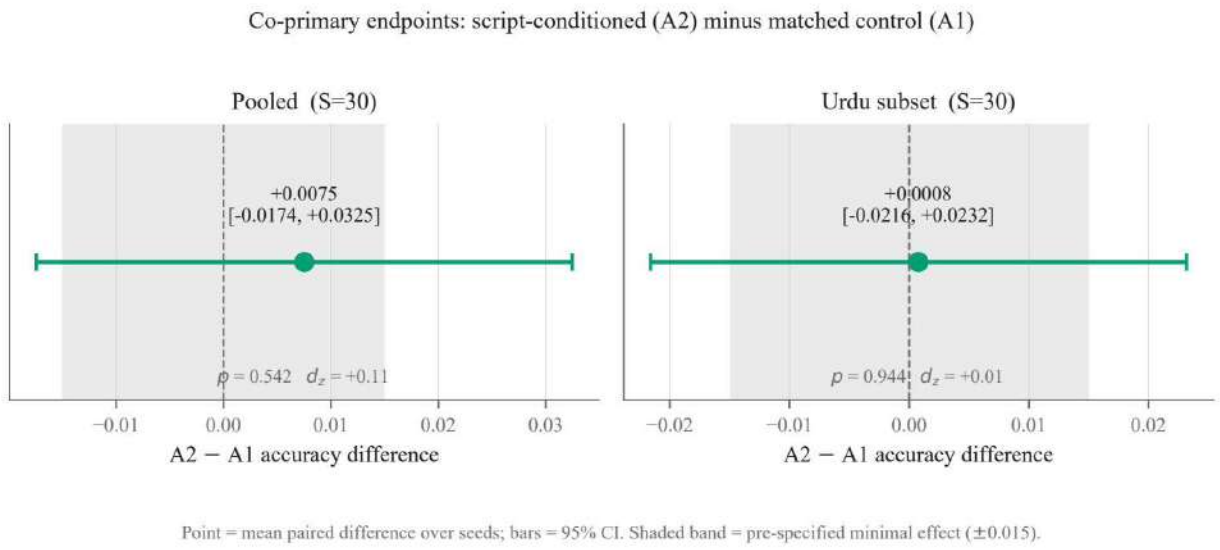

Figure 3. Co-primary endpoints. Point is the mean seed-paired difference over 30 seeds; whiskers are the 95% confidence interval, not a standard deviation. The shaded band spans the pre-specified minimal effect of interest, ± 0.015. Both intervals contain zero and both extend well beyond the band in each direction, which is the visual statement of the power limitation: this design cannot resolve effects of the size it was designed to look for.

How this must be read. The correct statement is that this design could not detect an effect of the pre-specified size. It is not a demonstration that script conditioning has no effect.

At an achieved power of 0.121, a true effect of 1.5 accuracy points would have been missed roughly seven times in eight. What the data do support is a bound: the pooled interval excludes an advantage for script conditioning larger than about 3.3 accuracy points, and the Urdu interval excludes one larger than about 2.3 points. Given that the study was motivated by the expectation of a gain concentrated in Urdu, the Urdu point estimate of +0.0008 is the more informative of the two, and it is as close to exactly zero as this measurement can resolve.

8. Results: Why the Null

A null result at low power invites the reading that the effect is real but unmeasured. We can constrain that reading directly, by measuring whether the conditioning signal does anything to the mechanism it was introduced to steer.

The probe. Because the module produces a channel gate and no spatial map (Section 5.2), there is no spatial attention to visualise, and producing a heat map by some other method and presenting it as the module's attention would misrepresent what the model computes. We use a counterfactual instead. For each of the 30 final seeds we reload the trained weights and recompute the gate vector for every test image under each of the three possible script identifiers, giving a 30 × 3 × 62 × 512 tensor of gate values. We then define two quantities. *Script-induced* variation is the standard deviation across the three forced script identifiers with the image held fixed, averaged over images and channels. *Image-induced* variation is the standard deviation across the 62 images with each image taking its true script identifier, averaged over channels.

The result. Script-induced gate variation is 2.82×10^{-4}. Image-induced variation is 8.61×10^{-3}. The ratio is **0.0327**: changing which script the model is told it is looking at moves the attention gates about 3% as much as changing the image does. The gates themselves are not inert, occupying a range from 0.172 to 0.841 about a mean of 0.499; they are simply almost unmoved by the script.

The probe reads the right tensor. The control arm provides a null check. Its gate computation does not reference the script identifier at all, so its script-induced variation must be indistinguishable from zero. Measured, it is 4.87×10^{-9}, float32 round-off, and roughly 58,000 times smaller than the treatment arm's. Any larger value would have meant the probe was reading the wrong quantity.

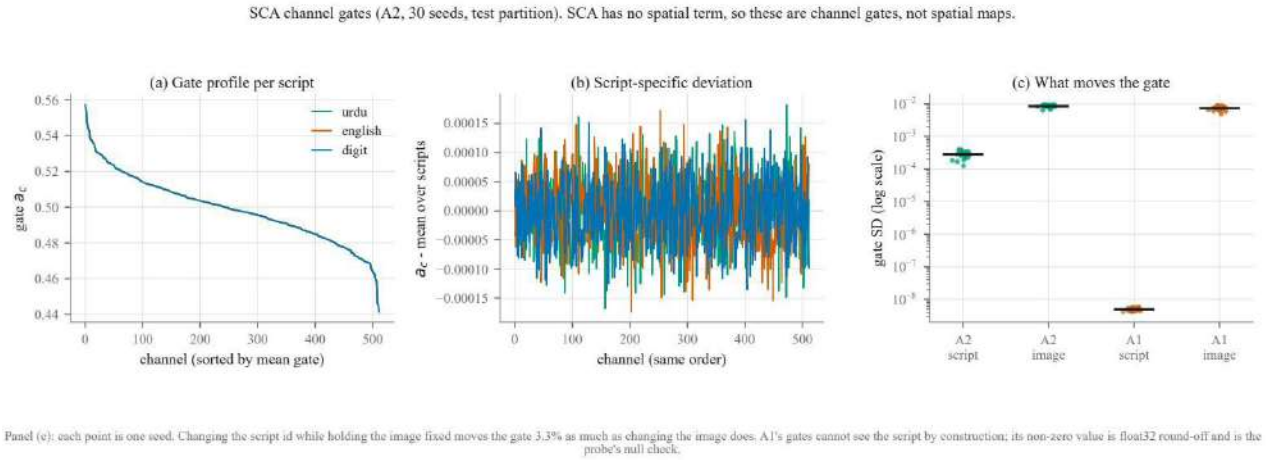

Figure 4. Channel-gate profiles and the script counterfactual. These are channel gates, not spatial attention maps: the module has no spatial component, so no spatial map exists to plot. Panel (a) shows the mean gate per channel under each forced script identifier; the three curves are visually superimposed, which is the finding. Panel (b) shows the per-script deviation from the cross-script mean, on an axis three orders of magnitude finer. Panel (c) compares script-induced against image-induced gate variation for both arms on a log scale, one point per seed, including the control arm's null check.

What this does and does not establish. It gives the null a mechanism: on this corpus, at this insertion depth, the trained module learned gates that are nearly invariant to the conditioning input, so there is little for the downstream classifier to exploit and the statistical null is what one would predict. It does not establish that script conditioning is unhelpful in general, that a different conditioning mechanism would fail, or that a spatial mechanism would fail. It localises *this* null to a measurable property of *this* trained module.

9. Results: Exploratory Context

Exploratory. Nothing in this section was pre-registered as a confirmatory endpoint. Values are means with dispersion. No significance is claimed and no *p*-value here carries a decision.

9.1 All arms

Table 7 reports every arm pooled and Table 5 per script. Three observations are worth recording, all descriptive.

The simplest baseline is not beaten. A0, the plain frozen backbone with the source study's head and no conditioning apparatus whatever, reaches 0.7554 mean accuracy, above both confirmatory arms. The 9.7-million-parameter conditioning apparatus buys nothing over it here. This observation inherits the epoch-selection bias of Section 6.4 and is offered as description, not as a ranking.

Under predicted script, conditioning is worse than not conditioning. A2′, the only arm driven by a predicted rather than an oracle script tag, reaches 0.7360, below the script-agnostic control's 0.7419. If script conditioning were to be deployed, this is the condition it would be deployed under. Its script classifier reaches 0.8559 ± 0.0239 accuracy on the test partition across the same 30 seeds, ranging from 0.8065 to 0.9032, so the degradation is produced by a classifier that is right

roughly six times in seven rather than by a poor one. We note a shortcoming in our own record-keeping: these values were written to a run log rather than to a structured artifact, and their per-script breakdown, which would show whether identification errors concentrate in one script, was never computed.

The routed arms are oracle-routed. A5 attains the highest pooled accuracy in the table at 0.7720, and A4 the lowest at 0.7199. Both route on the ground-truth script tag. A5 is therefore identify-then-branch with a perfect identifier, and its figure is not attainable without one. A4's digit accuracy is 0.5061, at chance, which is the expected consequence of sending every digit to an expert trained only on English characters.

Arm	*n*	Accuracy	Bal. acc.	Precision	Recall	F_1	AUC	Params
A0	30	0.7554 ± 0.0347	0.6737 ± 0.0575	0.7601 ± 0.0375	0.9268 ± 0.0351	0.8340 ± 0.0192	0.7582 ± 0.0439	9,651,649
A1	30	0.7419 ± 0.0536	0.6760 ± 0.0759	0.7711 ± 0.0528	0.8805 ± 0.0909	0.8174 ± 0.0438	0.7758 ± 0.0618	9,685,537
A2	30	0.7495 ± 0.0497	0.6878 ± 0.0769	0.7794 ± 0.0543	0.8789 ± 0.0700	0.8226 ± 0.0330	0.7779 ± 0.0521	9,685,537
A2′	30	0.7360 ± 0.0414	0.6645 ± 0.0678	0.7615 ± 0.0512	0.8862 ± 0.0629	0.8162 ± 0.0262	0.7794 ± 0.0500	9,685,537
A3	30	0.7253 ± 0.0560	0.6757 ± 0.0788	0.7790 ± 0.0600	0.8293 ± 0.0885	0.7988 ± 0.0437	0.7620 ± 0.0623	9,651,652
A4	30	0.7199 ± 0.0429	0.6794 ± 0.0402	0.7812 ± 0.0329	0.8049 ± 0.0859	0.7898 ± 0.0403	0.7588 ± 0.0340	19,303,298
A5	30	0.7720 ± 0.0390	0.7045 ± 0.0654	0.7850 ± 0.0511	0.9138 ± 0.0671	0.8413 ± 0.0258	0.7614 ± 0.1049	22,497,219

Table 7. Pooled test-set performance of every arm, mean ± standard deviation over 30 seeds on the fixed split. Absolute values are within-dataset and are not generalisation estimates. Arm A2′ appears as A2p in results/all_runs.csv; the label is identical. Generated from results/all_runs.csv.

Arm	Urdu (*n* = 43)	English (*n* = 8)	Digits (*n* = 11)
A0	0.7566 ± 0.0310	0.8333 ± 0.1327	0.6939 ± 0.0909
A1	0.7357 ± 0.0527	0.8667 ± 0.1502	0.6758 ± 0.0975
A2	0.7364 ± 0.0378	0.9125 ± 0.1319	0.6818 ± 0.1345
A2′	0.7341 ± 0.0360	0.9083 ± 0.1225	0.6182 ± 0.1227
A3	0.7202 ± 0.0670	0.7708 ± 0.1232	0.7121 ± 0.0864
A4	0.7636 ± 0.0536	0.7792 ± 0.1341	0.5061 ± 0.0780
A5	0.7574 ± 0.0413	0.8417 ± 0.1639	0.7788 ± 0.1086

Table 5. Exploratory. Per-script test accuracy, mean ± SD over 30 seeds. No significance is claimed for any cell. The English and digit columns rest on 8 and 11 images respectively, where a single image is worth 12.5 and 9.1 accuracy points; they are reported for completeness and are not interpretable. Generated from results/all_runs.csv.

9.2 The source study's backbones, re-run

Table 8 re-runs the source study's backbones under our single fixed split, so that comparisons within this paper are internally consistent rather than borrowed across differing partitions. Two results require comment.

Two backbones did not train. A convolutional network trained from scratch and MobileNetV3-Small both produced balanced accuracy of exactly 0.5000 with recall 1.000 and zero variance across every seed and learning rate: they emit the positive class for every image. Their apparent accuracy of 0.6613 is the base rate, not performance. Their AUC separates the two failures: 0.5029 for the from-scratch network, which is chance, against 0.5482 for MobileNetV3-Small, whose frozen features carry a little ranking signal that its head never converts into a decision. For the from-scratch network this is the expected consequence of the data budget: 618 images with no augmentation is not enough to train a convolutional network from random initialisation. We report both rows rather than dropping them, flagged, because reporting the base rate beside genuine numbers would imply they are weak-but-working models. They are not models in any useful sense.

Our MobileNetV1 anchor does not lead the table. Corrected, the anchor reaches 0.7258 validation accuracy with a balanced accuracy of 0.6378: joint-lowest on accuracy with InceptionV3 and lowest on balanced accuracy among the four non-degenerate backbones. VGG16 and MobileNetV2 both score above it. We record this because it runs against the source study's finding and against our own expectation, and because it was concealed until recently by a defect in our table generator that pooled the anchor's runs with those of six other architectures.

Table 8 deliberately carries no column comparing these figures with the source study's own. Such a column would require transcribing their results table, and we could not retrieve the published full text to transcribe it from. A transcription of it does exist in our working repository, made earlier and without a verification pass, and we have excluded it rather than reproduce a number whose only provenance is an unchecked reading. The table is therefore an internal-consistency result: our backbones, our split, compared with each other and with the base rate.

Nothing here should be read as a claim about the source study's results. Our split, preprocessing and training protocol all differ from theirs, and our deduplicated corpus is smaller than the file set they used, so the two are not commensurable even when both numbers are in hand.

Backbone	Val. acc.	Bal. acc.	AUC	Recall
mobilenet	0.7258 ± 0.0323	0.6378	0.7902	0.911
cnn_scratch†	0.6613 ± 0.0000	0.5000	0.5029	1.000
vgg16	0.7473 ± 0.0406	0.6928	0.7902	0.862
inceptionv3	0.7258 ± 0.0279	0.7153	0.7944	0.748
mobilenetv2	0.7419 ± 0.0161	0.6887	0.7530	0.854
mobilenetv3small†	0.6613 ± 0.0000	0.5000	0.5482	1.000

Table 8. A0 backbone comparison on the validation partition, mean ± SD over 3 seeds at each backbone's best learning rate. Daggered rows collapsed to a constant majority-class predictor (balanced accuracy 0.5, recall 1.0); their accuracy is the base rate 0.6613, not performance. This is an internal-consistency comparison under one fixed split; it is not set against the source study, whose full text was not retrievable.

9.3 Conditioning depth, and naive conditioning

Across the four insertion depths the validation difference between the conditioned arm and its control was 0.0000, −0.0054, −0.0108 and −0.0108: no advantage at any depth, and a consistent null-to-negative sign. Against naive conditioning, which concatenates a one-hot script vector onto the penultimate features and uses no attention module at all, the conditioned arm's pooled difference is +0.0242 with a 95% interval

of [−0.0075, +0.0559]. That interval contains zero, and no significance is claimed for it.

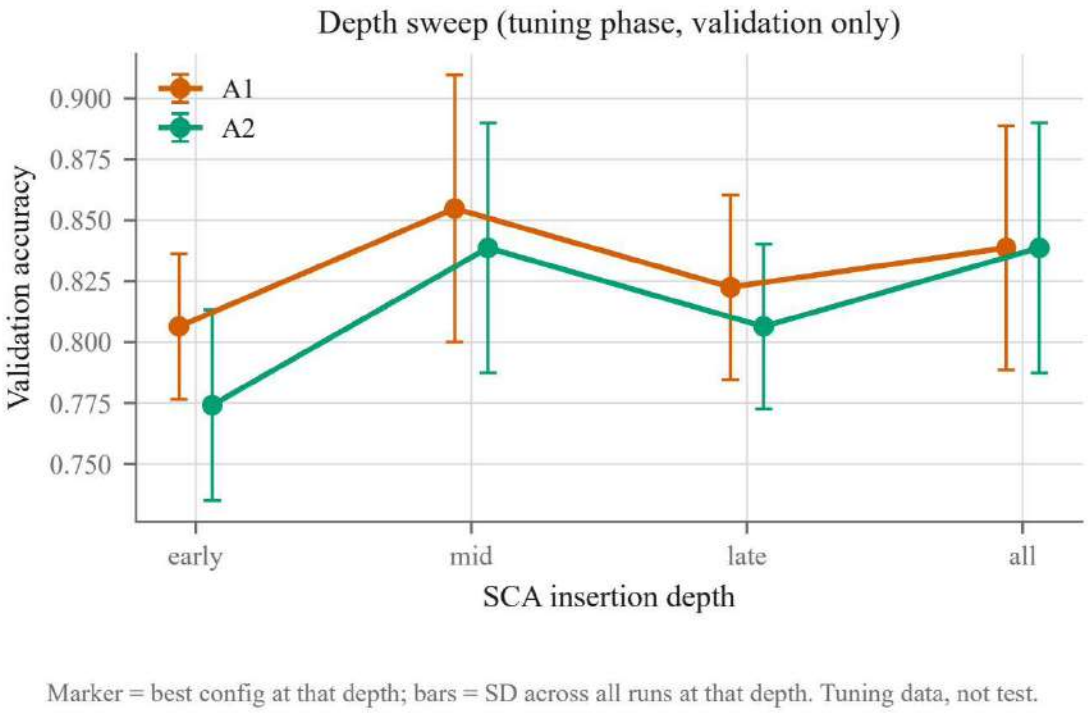

Figure 5. Conditioning depth, validation partition, tuning phase only. Note the plotted quantities: the marker is the best single configuration at that depth and the bar is the standard deviation across all runs at that depth, so the bar is not an uncertainty interval on the marker.

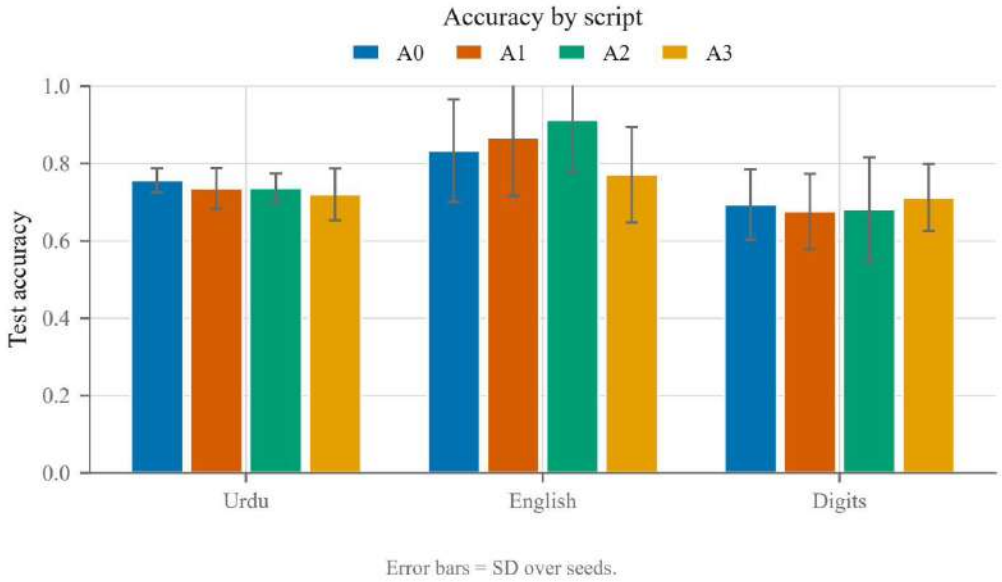

Figure 6. Accuracy by script for the four principal arms; error bars are the standard deviation over seeds. The English and digit groups rest on 8 and 11 test images respectively and are not interpretable at that size; they are shown for completeness.

10. Discussion

The result of this study is a bound, not a verdict. Script conditioning, implemented as a capacity-matched channel-attention module and evaluated under a pre-registered protocol on this corpus, does not produce a detectable improvement, and the design can exclude improvements larger than roughly three accuracy points. It cannot exclude the effect it was built to look for. Anyone reading this paper as evidence that script conditioning does not work is reading more than the data support.

What raises the null above an inconclusive report is the mechanism. The gate measurement shows that the conditioning input barely moves the attention it steers, which supplies a specific and testable reason for the statistical result. It also suggests where the next attempt should differ. The module conditions a channel gate, and channel gates apparently learn to be nearly script-invariant on this corpus; a mechanism with a spatial term, or conditioning applied earlier in the network where script-specific stroke statistics are still spatially localised, would be a different experiment rather than a repetition of this one.

The audit generalises further than the ablation does. Its specific finding concerns one release, but the procedure costs a single pass of checksums and it changed this study's corpus, its class balance, its baseline and its split design. That a corpus can pass a file-count gate and fail a content check is not a property of this dataset in particular. The reporting convention we would propose is minimal: publish distinct-content counts alongside file counts, report cross-class collisions explicitly, and state whether the deduplication is exact or perceptual.

Finally, the honest deployment number in this paper is not the confirmatory contrast. It is A2′, the only arm driven by a predicted script, which lands below the script-agnostic control. Under conditions resembling use, adding script conditioning to this pipeline made it slightly worse.

11. Limitations

1. **Within-dataset scope only.** Every figure is within-dataset separability on one corpus. None is a generalisation estimate, and no screening or diagnostic claim is made or supported.
2. **Site-and-class confound.** The dyslexic and non-dyslexic samples were collected at different institutions, so the class label is confounded with collection site. Both arms face the confound identically, which is why a matched-pair contrast remains interpretable while absolute accuracy does not.
3. **Construct validity of the labels.** The label is institutional placement supported by psychological assessment, not an independent clinical dyslexia diagnosis, and we were unable to verify the selection procedure directly because the source study's full text was not accessible to us. Construct validity therefore rests on a labelling process we have not been able to inspect. This is a limitation of our position as secondary analysts, not a criticism of the original collection.
4. **Writer-disjoint evaluation is impossible.** The release contains no writer identifiers and no per-writer grouping, so images from the same child may fall on both sides of the partition. This is the most serious limitation of the corpus and it cannot be repaired by any choice of split. [4] demonstrate what is possible when participant identifiers exist; that option is not available here.
5. **The equal-separability assumption.** The matched-pair logic assumes the confounds above are equally exploitable by both arms. We argue this is reasonable, since the arms are identical in size and architecture and differ only in one input, but we cannot prove it.
6. **The audit is a lower bound.** Deduplication is exact-hash. Near-duplicates that were re-encoded or rescaled survive it, and a perceptual scan proved unreliable on these images and was discarded. Residual near-duplicate leakage cannot be excluded. Writer-level leakage cannot even be detected from the released metadata.
7. **Annotation provenance.** The agreement statistic ($\kappa = 0.8489$) is human-versus-AI, not inter-human. Adjudication was human-authoritative by fixed policy rather than case-by-case joint

resolution, and the released file's ambiguity column is the AI annotator's throughout. Script tags are a covariate rather than the outcome label, so annotation noise affects both confirmatory arms identically and does not bias their contrast, but the released layer should be used with its provenance in view.

8. **Power.** Achieved power is 0.121 for the pre-specified effect. The smallest detectable effect at 80% power is 4.33 accuracy points. The test partition holds 62 images, so one image is worth 1.61 points and the pre-specified effect sits below the resolution of a single-seed measurement. This is the binding limitation on the confirmatory result.

9. **Epoch selection on the evaluation partition.** As disclosed in Section 6.4, six of seven arms early-stopped and restored weights on the test partition. Absolute figures are optimistically biased, and A4, the single exception, is not comparable with the rest.

10. **Per-script subset sizes.** The test partition holds 43 Urdu, 8 English and 11 digit images. The Urdu co-primary endpoint rests on 43 images. The English and digit columns are reported with four decimals for consistency of format only and should not be interpreted.

11. **A pre-registered analysis was not run.** The sensitivity corpus of 852 released files was defined and split but never evaluated.

12. Conclusion

A published bilingual dyslexia handwriting corpus contains 638 distinct images in 852 files, with 20 images carrying contradictory labels across both diagnostic classes and a stated class balance that does not survive deduplication. On the corrected corpus, a pre-registered, capacity-matched, 30-seed ablation finds no evidence that script-conditioned channel attention improves binary dyslexia classification over an identical-size script-agnostic control, and a counterfactual probe shows why: the script input moves the attention gates about 3% as much as the image does. The design's achieved power is 0.121 for its pre-specified effect, so the null bounds the effect rather than refuting it. We regard the audit as the finding most likely to be useful elsewhere, and we would encourage the simple reporting convention it implies: publish content counts, not only file counts.

Declarations

Data availability. We do not redistribute the handwriting images; they remain available from [5] through the access route published there. Our own artifacts, namely the code, configuration files, script annotations, fixed split indices, per-seed raw results and figure-generation scripts, are available at https://github.com/AyeshaAshfaq12/Script-Aware-Bilingual-Dyslexia-Classification. The repository includes the frozen configuration files, the freeze-v1 tag that gates test-set evaluation, and the per-seed run log from which every table and figure in this paper is generated.

Ethics. This study is a secondary analysis of a corpus released openly by its original authors under CC BY 4.0. It involved no new data collection, no contact with participants and no attempt to identify any individual. We do not redistribute the images. Because the source article's full text was not accessible to us, we make no representation about its consent, assent or institutional approval record; those are documented in that article and readers should consult it there.

Author contributions. Following the CRediT taxonomy. **Ayesha Ashfaq:** conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing (original draft). **Abdullah Butt:** conceptualization, methodology, supervision, project administration, writing (review and editing). Both authors read and approved the final manuscript and accept responsibility for its contents.

Conflicts of interest. The authors declare no conflicts of interest.

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

Use of AI tools. Development of the companion code and drafting of this manuscript used Claude

and Claude Code. An AI assistant additionally served as annotator A in the script-annotation pass; this is disclosed in Section 4.2 and is the reason the agreement statistic is reported as human-versus-AI rather than inter-human. All script tags in the released annotation file were adjudicated by a human author, who is accountable for them. The authors take full responsibility for the content of this manuscript, including every claim, citation and reported value, and no AI system is credited as an author.

REFERENCES

[1] P. Suárez-Coalla, A. Carnota Orviz, M. González-Nosti, and C. Martínez-García. Writing performance in Spanish adults with dyslexia: Handwriting versus typing. *Reading and Writing*, 39:529–558, 2026. doi: 10.1007/s11145-025-10709-w.

[2] J. Kunhoth, S. Al-Maadeed, S. Kunhoth, and Y. Akbari. Automated systems for diagnosis of dysgraphia in children: A survey and novel framework, 2022. doi: 10.48550/arXiv.2206.13043. Preprint; no journal reference recorded on arXiv.

[3] M. Robaa, M. Balat, R. Awaad, E. Omar, and S. A. Aly. Explainable AI in handwriting detection for dyslexia using transfer learning, 2024. Preprint; journal status not established.

[4] H. W. Liu, S. Wang, and S. X. Tong. DysDiTect: Dyslexia identification using CNN-positional-LSTM-attention modeling with Chinese dictation task. *Brain Sciences*, 14(5):444, 2024. doi: 10.3390/brainsci14050444.

[5] M. Kashif, Z. M. Haider, I. Muneer, D. Kumar, T. Tahir, and J. Shafi. Optimizing deep neural models for early dyslexia detection using novel bilingual handwritten dataset. *Engineering Reports*, 8(6):e70867, 2026. doi: 10.1002/eng2.70867. Open access under CC BY 4.0.

[6] J. Mekyska, Z. Galaz, K. Safarova, V. Zvoncak, L. Cunek, T. Urbanek, J. M. Havigerova, J. Bednarova, J. Mucha, M. Gavenciak, Z. Smekal, and M. Faundez-Zanuy. Assessment of developmental dysgraphia utilising a display tablet. In *Graphonomics in Human Body Movement. Bridging Research and Practice from Motor Control to Handwriting Analysis and Recognition*, volume 14285 of *Lecture Notes in Computer Science*, pages 21–35. Springer, 2023. doi: 10.1007/978-3-031-45461-5_2.

[7] M. Bublin, F. Werner, A. Kerschbaumer, G. Korak, S. Geyer, L. Rettinger, E. Schönthaler, and M. Schmid-Kietreiber. Automated dysgraphia detection by deep learning with SensoGrip, 2022. doi: 10.48550/arXiv.2210.07659. Preprint; listed on DBLP with no peer-reviewed venue.

[8] H. Maryam, L. Fu, J. Song, T. A. B. M. Shafayet, Q. Luo, X. Bai, and Y. Liu. Dataset and benchmark for Urdu natural scenes text detection, recognition and visual question answering. In *Document Analysis and Recognition – ICDAR 2024*, volume 14808 of *Lecture Notes in Computer Science*, pages 279–292. Springer, 2024. doi: 10.1007/978-3-031-70549-6_17.

[9] S. Ihbour, M. Slimani, B. Mohamed, H. Chahbi, H. El Azzouzi, H. Anarghou, D. Rabih, K. Kaoutar, and F. Chigr. Cognitive assessment and progress in language skills in Arabic-speaking subjects with neurodevelopmental dyslexia. *Journal of Neurosciences in Rural Practice*, 17(Suppl. 1):S84, 2026. doi: 10.25259/JNRP_298_2025.

[10] Z. Chen, F. Yin, X.-Y. Zhang, Q. Yang, and C.-L. Liu. MuLTReNets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition. *Pattern Recognition*, 108:107555, 2020. doi: 10.1016/j.patcog.2020.107555.

[11] A. A. S. Azeez. Hybrid deep learning-crow search algorithm for accurate Arabic handwritten digit recognition. *Kufa Journal of Engineering*, 17(3):601–622, 2026. doi: 10.30572/2018/KJE/170334.

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671.