THE GREEDY SUPERSTRING RATIO AT STRING LENGTHS FOUR AND FIVE: $\rho_4 > 15/8$ AND $\rho_5 > 15/8$

KOREA SUPERINTELLIGENCE LABS

ABSTRACT. Given a finite set S of strings, GREEDY repeatedly replaces a pair of distinct elements of maximum overlap by their overlap merge; ties are broken arbitrarily, and ρ_k denotes the supremum of $|g|/\mathrm{OPT}(S)$ over all instances $S \subseteq \Sigma^k$ and all outputs g of a legal execution. Cazaux and Rivals proved $2 - 1/k \le \rho_k \le \min\{(k+1)/2,\ 7/2\}$ and conjectured that the lower bound is the exact value. Chukhin, Kulikov, Mihajlin and Smal have recently refuted that conjecture at $k = 3$, where they determine $\rho_3 = 9/5$, and at every $k \ge 6$, where they prove $\rho_k \ge 2$; the two string lengths $k = 4$ and $k = 5$ are reached by neither of their results and are recorded as open in their introduction. We refute the conjecture at both, without determining either value. We exhibit an instance of 230 strings of length 4 whose shortest common superstring has length exactly 233 and on which one legal greedy execution outputs a string of length 437, and an instance of 169 strings of length 5 with optimum exactly 173 and a legal greedy output of length 325; hence $\rho_4 \ge 437/233 > 15/8$ and $\rho_5 \ge 325/173 > 15/8$, against the conjectured values $7/4$ and $9/5$. In particular $\rho_4 > 9/5 = \rho_3$: greedy is strictly worse on strings of length 4 than on strings of length 3, which the previously known bracket $1.75 \le \rho_4 \le 2$ did not decide. With the results quoted above, the value $2-1/k$ is now known to be wrong at every $k \ge 3$. In each instance the optimum is proved by a counting bound together with an exhibited witness, so no search over superstrings is involved; the executions are certified merge by merge. Every statement about an individual instance below — the counting bound, the optimality of each exhibited witness, and the legality of each execution — has been checked by the Lean 4 kernel; the final section says exactly what that does and does not cover.

1. INTRODUCTION

1.1. **Superstrings and the greedy algorithm.** Let Σ be a finite alphabet. A string w is a *common superstring* of a finite set S of strings if every element of S occurs in w as a contiguous block, and $\mathrm{OPT}(S)$ denotes the length of a shortest common superstring. Computing OPT is NP-hard, and the algorithm that has dominated the subject for forty years is the simplest one available. For strings p, q write

$$\mathrm{ov}(p,q) \;=\; \max\{\,\ell \;:\; \text{the length-}\ell \text{ suffix of } p \text{ equals the length-}\ell \text{ prefix of } q\,\},$$

and let $p \odot q$ be the *overlap merge*, that is p followed by all of q except its first $\mathrm{ov}(p,q)$ letters. GREEDY is:

> while more than one string is live, choose an ordered pair of distinct live strings (p, q) with $\mathrm{ov}(p,q)$ maximum, and replace p and q by $p \odot q$.

Date: August 30, 2026.

Ties are broken arbitrarily, so GREEDY is a nondeterministic algorithm and GREEDY(S) is a *set* of outputs, each of them a common superstring of S. The *greedy conjecture* asserts that $|g| \le 2\,\mathrm{OPT}(S)$ for every S and every $g \in \textsc{Greedy}(S)$. It dates from the late 1980s. Blum, Jiang, Li, Tromp and Yannakakis [6] state it in the form used here — a common conjecture, in their words, that GREEDY "produces a superstring of length at most two times optimal" — and attribute it there to Storer [7], to Tarhio and Ukkonen [8] and to Turner [9]; those same four sources are the ones [1] cites for it. The conjecture is open.

For $k \ge 2$, *k-SCS* is the restriction of the problem to instances $S \subseteq \Sigma^k$, all input strings having the same length k, and

$$\rho_k \;=\; \sup\Big\{\, \frac{|g|}{\mathrm{OPT}(S)} \;:\; S \subseteq \Sigma^k \text{ finite},\ g \in \textsc{Greedy}(S) \,\Big\}. \tag{1}$$

The greedy conjecture restricted to k-SCS says $\rho_k \le 2$. Two features of (1) matter throughout: the supremum is over *all* legal executions, so a lower bound on ρ_k is witnessed by a single instance together with a single adversarial tie-breaking; and inputs of a common length are automatically substring-free, so the variant of GREEDY that first discards strings contained in others begins with the same instance — that no live string is ever contained in another along the executions exhibited below is recorded in Section 9.

1.2. **The Cazaux–Rivals conjecture and the two open cells.** Cazaux and Rivals [2] proved

$$2 - \frac{1}{k} \;\le\; \rho_k \;\le\; \min\Big\{\frac{k+1}{2},\ \frac{7}{2}\Big\}, \tag{2}$$

the lower bound by an explicit series of instances (their Proposition 2) and the upper bound as their Theorem 3, and immediately after it they wrote:

> "Theorem 3 suggests a more precise version of the greedy conjecture: the superstring approximation ratio of greedy on r-SSP is $2 - 1/r$."

We refer to the assertion $\rho_k = 2 - 1/k$ as the Cazaux–Rivals conjecture.

Chukhin, Kulikov, Mihajlin and Smal [1] refuted it at both ends of the range. They prove $\rho_k \ge 2$ for every $k \ge 6$, so that the conjectured value is wrong there by a whole $1/k$, and they determine $\rho_3 = 9/5$ exactly, against the conjectured $5/3$. Their introduction records the resulting state of the art as

$$\begin{gathered} 1.66 \le \rho_3 \le 2, \qquad 1.75 \le \rho_4 \le 2.5, \qquad 1.8 \le \rho_5 \le 3, \\ 2 - \tfrac{1}{k} \le \rho_k \le 3.396 \quad (k \ge 6), \end{gathered} \tag{3}$$

the lower entries being the Cazaux–Rivals values $2 - 1/k$, with $5/3$ printed as 1.66. Neither of their two theorems reaches $k = 4$ or $k = 5$, and the abstract of the current version of [1] claims a complete characterisation only "for strings of length 3". Those two string lengths are the subject of this paper.

Remark 1. The entry $\rho_4 \le 2.5$ of (3) is the generic bound $(k+1)/2$ of (2). In fact $\rho_4 \le 2$ has been known since Kulikov, Savinov and Sluzhaev [3], whose abstract reads, in full, "In this short note, we prove that the greedy conjecture for the shortest common superstring problem is true for strings of length 4"; Cazaux and Rivals report the same result as "in the case where all input words have length 4 (for 4-SSP) the greedy algorithm achieves a superstring ratio of at most 2" [2],

and Nikolaev's survey lists it as "GC holds for strings of length exactly 4" [4, v1, §1]. So the sharpest published bracket at $k = 4$ is [1.75, 2] rather than [1.75, 2.5]. This is not an error in anything [1] proves; we note it because it is the bracket our results should be measured against, and because it makes the comparison $\rho_4 > \rho_3$ of Corollary 4 meaningful — the published bracket is consistent with $\rho_4 = \rho_3 = 9/5$. We have not been able to consult the text of [3], which is closed access with no open copy, and say nothing here about how it is proved.

1.3. **Results.** Everything below concerns explicit instances. Two block families are constructed in Section 3, one for each of the two string lengths; $S_t^{(4)}$ consists of $10t$ strings of length 4 and $S_t^{(5)}$ of $13t$ strings of length 5. The main theorems are the following four statements, each an instance of a single scheme: an instance, one legal greedy execution on it, and a common superstring proved to be shortest.

Theorem 2 (Main theorem)**.**
(i) $\rho_4 \geq 76/43 > 7/4$, *witnessed by* $S_4^{(4)}$ *(*40 *strings, optimum* 43*, a legal greedy output of length* 76*).*
(ii) $\rho_5 \geq 125/69 > 9/5$, *witnessed by* $S_5^{(5)}$ *(*65 *strings, optimum* 69*, a legal greedy output of length* 125*).*
(iii) $\rho_4 \geq 437/233 > 15/8$, *witnessed by* $S_{23}^{(4)}$ *(*230 *strings, optimum* 233*, a legal greedy output of length* 437*); already* $S_6^{(4)}$ *gives* $\rho_4 \geq 114/63 > 9/5$.
(iv) $\rho_5 \geq 325/173 > 15/8$, *witnessed by* $S_{13}^{(5)}$ *(*169 *strings, optimum* 173*, a legal greedy output of length* 325*); already* $S_7^{(5)}$ *gives* $\rho_5 \geq 175/95 > 11/6$.

Parts (i) and (ii) refute the Cazaux–Rivals conjecture at the two string lengths it survived; parts (iii) and (iv) say by how much. Three consequences are worth stating separately.

Corollary 3. *The Cazaux–Rivals conjecture* $\rho_k = 2 - 1/k$ *is false for every* $k \geq 3$.

Corollary 4. $\rho_4 > 9/5 = \rho_3$*: the greedy superstring ratio is strictly larger on strings of length* 4 *than on strings of length* 3.

Corollary 5. $15/8 < \rho_4 \leq 2$ *and* $15/8 < \rho_5 \leq 3$.

Corollary 3 combines Theorem 2 at $k = 4, 5$ with the two theorems of [1] at $k = 3$ and $k \geq 6$; at $k = 3$ and $k = 6$ their own constructions are reproduced here as finite instances, checked without appeal to their proofs (Theorem 15(iii)), so the only case that rests on a quoted theorem is $k \geq 7$. Corollary 4 combines Theorem 2(iii) with the equality $\rho_3 = 9/5$ of [1], which we quote and do not reprove; the inequality $\rho_4 > 9/5$ itself is proved here. Corollary 5 adds the upper bounds of [3] (see Remark 1) and of (2); no upper bound is proved in this paper.

Theorem 2 is Theorems 10 and 11 at $k = 4$ and Theorems 12 and 13 at $k = 5$; the corollaries are proved in Section 6. The two families continue past the instances listed above; what they converge to is described in Remark 9, which is not part of the machine-checked development.

1.4. **Method.** Each of the four statements of Theorem 2 needs two numbers about one explicit instance S: an upper bound on $\mathrm{OPT}(S)$ and a lower bound on the length of some legal greedy output.

The optimum is obtained without any search. Exhibiting a common superstring w of S bounds $\mathrm{OPT}(S) \leq |w|$, and that inequality is all the ratio needs; but each of

our instances comes with a witness of length exactly $n+k-1$, which by Lemma 7 — an elementary counting argument — is the least a common superstring of n distinct strings of length k can have. So $\mathrm{OPT}(S)$ is determined exactly, and the ratio reported is the instance's true ratio and not merely a lower bound for it. The one place where a mistake would flatter the result is the claim that the exhibited word is a common superstring at all: were it not, the true optimum could be larger and the ratio smaller. That claim is a decidable property of explicit data, it is checked, and Theorem 16(iv) shows the check is not vacuous.

The greedy output is a certificate: a list of merges, each recorded as an ordered pair of positions in the current list of live strings. Replaying it is a finite computation which checks, at every step, that the two chosen live strings are distinct and that their overlap equals the maximum overlap over *all* ordered pairs of distinct live strings — so the certificate is accepted only if every merge is one Greedy is permitted to make. The replay is the largest computation entering any theorem below: at each of the 229 steps of the longest certificate the checker maximises over up to 230^2 ordered pairs.

How the two block words were found is a separate matter, described in Section 8. That section reports a search, but no theorem in this paper depends on it: the blocks are simply data once written down.

2. Preliminaries

2.1. **Strings, overlap, merge.** Fix an alphabet Σ. Strings are finite sequences over Σ, $|w|$ is the length of w, and $s \sqsubseteq w$ means that s occurs in w as a contiguous block, that is $w = u\,s\,v$ for some strings u, v. For a finite set S of strings,

$$\mathrm{OPT}(S) \;=\; \min\{\, |w| \;:\; s \sqsubseteq w \text{ for all } s \in S \,\}.$$

For strings p, q the overlap $\mathrm{ov}(p,q)$ is the largest $\ell \le \min(|p|,|q|)$ such that the length-ℓ suffix of p equals the length-ℓ prefix of q; it always exists, $\ell = 0$ being admissible. The overlap merge is $p \odot q = p \cdot q[\mathrm{ov}(p,q) :]$, of length $|p|+|q|-\mathrm{ov}(p,q)$.

2.2. **Legal executions.** A *state* is a finite list L of live strings. Write

$$\mathrm{maxov}(L) \;=\; \max\{\, \mathrm{ov}(p,q) \;:\; p,q \in L,\; p \neq q \,\}$$

(the maximum over ordered pairs of *distinct strings*, with $\mathrm{maxov}(L) = 0$ when no such pair exists). A step is a pair of positions (i,j) of L; it is *legal* if $L_i \neq L_j$ and $\mathrm{ov}(L_i, L_j) = \mathrm{maxov}(L)$, and it transforms L into the list obtained by deleting positions i and j and appending $L_i \odot L_j$.

Definition 6. A *greedy certificate* for an instance S (listed in some order) is a sequence of steps, each legal in the state it is applied to, that reduces S to a single string g. Then $g \in \textsc{Greedy}(S)$, and g is a common superstring of S.

Definition 6 is deliberately conservative in two ways. It compares strings, not positions: if a state ever held two equal strings, a positions-only reading would allow a merge Greedy does not. And it forbids merging two occurrences of the same string, which is what "a pair of distinct strings" means. Neither restriction is ever active in the executions below, but each makes the certified executions a subset, never a superset, of the executions Greedy may perform.

2.3. **The counting bound.**

Lemma 7. *Let S be a nonempty finite set of pairwise distinct strings, each of length k, and let w be a common superstring of S. Then*

$$|w| \;\geq\; |S| + k - 1.$$

Proof. A string of length ℓ has exactly $\ell - k + 1$ starting positions for a block of length k (and none at all if $\ell < k$; since $S \neq \emptyset$, some $s \in S$ occurs in w, so $|w| \geq k$). Every $s \in S$ occurs at some starting position, distinct elements of S occupy distinct positions because a position determines its block, and hence $|S| \leq |w| - k + 1$. □

This is folklore; it is the elementary half of the lemma of Pevzner [5] quoted as such in [1], whose other half says that equality holds whenever the de Bruijn subgraph of S is balanced and weakly connected. Only the inequality is used here, and it is used in the form:

Lemma 8. *If in addition w is a common superstring of S with $|w| = |S| + k - 1$, then w is shortest, i.e.* $\mathrm{OPT}(S) = |S| + k - 1$.

Proof. Immediate from Lemma 7. □

2.4. **What a lower bound on ρ_k requires.** Every theorem in Sections 4–7 is an instance of the following statement, which we abbreviate $\rho_k > a/b$:

> there are a nonempty finite set S of pairwise distinct strings of length k, a greedy certificate for S with output g, and a common superstring w of S such that $|w| \leq |v|$ for every common superstring v of S, with $a\,|w| < b\,|g|$.

Given such data, $|g|/\mathrm{OPT}(S) = |g|/|w| > a/b$, and $\rho_k \geq |g|/\mathrm{OPT}(S)$ by (1), so indeed $\rho_k > a/b$. We emphasise that the formal development does not define ρ_k — it defines no supremum at all — and therefore asserts nothing beyond what one instance witnesses; the passage from the displayed statement to $\rho_k > a/b$ is the definition (1) and is carried out in prose, here and nowhere else.

3. The two block families

The instances of Theorems 10–13 are the members of two families described in this section. The description explains how the instances are built and why they behave as they do; it is not what the theorems rest on. Each theorem is proved on its own instance, whose optimum comes from Lemma 8 applied to an exhibited witness and whose execution is replayed merge by merge, so no property claimed below for a general t enters any proof.

3.1. **Blocks and spectra.** Let B be a cyclic word of length b over the alphabet $\{\mathtt{A}, \mathtt{B}, \mathtt{X}\}$, in which $\mathtt{A}$ and $\mathtt{B}$ are *shared* letters and $\mathtt{X}$ is *fresh*. For $t \geq 1$ let B_i denote B with $\mathtt{X}$ renamed to a letter $\mathtt{X}_i$ private to the index i, and let

$$S_t(B) \;=\; \bigcup_{i=1}^{t} \{\,\text{the } b \text{ cyclic } k\text{-mers of } B_i\,\},$$

where the cyclic k-mers of a cyclic word of length b are the b windows of length k read with wraparound. If the b cyclic k-mers of B are pairwise distinct and each contains $\mathtt{X}$, then $S_t(B)$ consists of exactly bt pairwise distinct strings of length k: distinct within a block by the first condition, and across blocks because a string of block i contains $\mathtt{X}_i$.

3.2. **The block at $k = 4$.** Take $k = 4$ and

$$B^{(4)} \;=\; \texttt{AAXABAXBAX} \qquad (b = 10).$$

Its ten cyclic 4-mers, read from positions $0, \ldots, 9$, are

$$\texttt{AAXA},\ \texttt{AXAB},\ \texttt{XABA},\ \texttt{ABAX},\ \texttt{BAXB},\ \texttt{AXBA},\ \texttt{XBAX},\ \texttt{BAXA},\ \texttt{AXAA},\ \texttt{XAAX},$$

pairwise distinct, each containing $\texttt{X}$. Write $S_t^{(4)} = S_t(B^{(4)})$, so $|S_t^{(4)}| = 10t$.

The optimum. Let $v^* = \texttt{ABA}$ (the unique cyclic 3-mer of $B^{(4)}$ free of $\texttt{X}$) and let B_i^{rot} be B_i rotated to begin at v^*, i.e. $\texttt{ABAX}_i\texttt{BAX}_i\texttt{AAX}_i$. Then

$$w_t^{(4)} \;=\; B_1^{\mathrm{rot}} B_2^{\mathrm{rot}} \cdots B_t^{\mathrm{rot}}\, v^* \tag{4}$$

has length $10t + 3$ and is a common superstring of $S_t^{(4)}$ — it traverses each block's cyclic word once, returning each time to v^* — so by Lemma 8, $\mathrm{OPT}(S_t^{(4)}) = 10t+3$.

The execution. The certificate makes greedy assemble, inside each block, the three strings

$$\texttt{X}_i\texttt{AAX}_i\texttt{AA}\ (6), \qquad \texttt{X}_i\texttt{ABAX}_i\texttt{AB}\ (7), \qquad \texttt{X}_i\texttt{BAX}_i\texttt{BA}\ (6), \tag{5}$$

of total length 19 per block, and then concatenates the resulting $3t$ strings with no further overlap, for an output of length $19t$. Two facts make every merge legal. First, each of the merges producing (5) joins two live strings at overlap $3 = k-1$, and by the endpoint-inheritance lemma of [6], recalled in Remark 17, no two distinct live strings can overlap in more than $k - 1$ letters. So these merges are legal whatever else is live. Second, once (5) is reached, no two of the $3t$ strings overlap at all. Indeed each of them begins and ends with one of the three words $\texttt{X}_i\texttt{AA}$, $\texttt{X}_i\texttt{AB}$, $\texttt{X}_i\texttt{BA}$, and any overlap of two of them is an overlap of length ≤ 3 between such words; for length 1 or 2 the suffix begins with a shared letter while the prefix begins with a fresh one, and for length 3 the two words would have to be equal, which forces the same block and the same string. Hence greedy's last $3t - 1$ merges are plain concatenations and

$$|g| \;=\; 19t, \qquad \mathrm{OPT}(S_t^{(4)}) \;=\; 10t + 3.$$

For the three values of t used in Section 4 both quantities are verified on the instance itself; for general t see Remark 9.

3.3. **The block at $k = 5$.** Take $k = 5$ and

$$B^{(5)} \;=\; \texttt{AAABXAABXBXBX} \qquad (b = 13),$$

whose thirteen cyclic 5-mers are pairwise distinct and each contain $\texttt{X}$; write $S_t^{(5)} = S_t(B^{(5)})$, so $|S_t^{(5)}| = 13t$. The $\texttt{X}$-free cyclic 4-mer is $v^* = \texttt{AAAB}$, the block already begins there, and

$$w_t^{(5)} \;=\; B_1 B_2 \cdots B_t\, v^* \tag{6}$$

is a common superstring of length $13t+4 = |S_t^{(5)}|+k-1$; by Lemma 8 it is shortest. The certificate assembles, per block,

$$\texttt{X}_i\texttt{AAABX}_i\texttt{AAA}\ (9), \qquad \texttt{X}_i\texttt{AABX}_i\texttt{BX}_i\texttt{AAB}\ (10), \qquad \texttt{X}_i\texttt{BX}_i\texttt{BX}_i\texttt{B}\ (6), \tag{7}$$

of total length 25, every merge again being at the maximum possible overlap $4 = k-1$, and then concatenates the $3t$ strings with no overlap, so that $|g| = 25t$ against $\mathrm{OPT}(S_t^{(5)}) = 13t + 4$. Here the pairwise-overlap-0 condition is not settled by the one-line argument of Section 3.2, because the third string of (7) begins with $\texttt{X}_i\texttt{BX}_i\texttt{B}$,

whose fresh letter is not only initial; it is instead a finite check over all ordered pairs of the three initial 4-mers, same-block and cross-block, which two blocks already exhaust. Again, for the three values of t used in Section 5 the two quantities are verified on the instance itself.

3.4. **The instances used below.** All six instances of Theorem 2 are members of these two families:

k	instance	n	OPT	$\lvert g\rvert$	ratio
4	$S_4^{(4)}$	40	43	76	$76/43 = 1.7674\ldots > 7/4$
4	$S_6^{(4)}$	60	63	114	$114/63 = 1.8095\ldots > 9/5$
4	$S_{23}^{(4)}$	230	233	437	$437/233 = 1.8755\ldots > 15/8$
5	$S_5^{(5)}$	65	69	125	$125/69 = 1.8115\ldots > 9/5$
5	$S_7^{(5)}$	91	95	175	$175/95 = 1.8421\ldots > 11/6$
5	$S_{13}^{(5)}$	169	173	325	$325/173 = 1.8786\ldots > 15/8$

Each t is the least at which $19t/(10t+3)$, respectively $25t/(13t+4)$, exceeds the value in the last column, so the margins at the threshold are narrow: the last two lines are $15 \cdot 233 = 3495 < 3496 = 8 \cdot 437$ and $15 \cdot 173 = 2595 < 2600 = 8 \cdot 325$.

Remark 9 (not part of the machine-checked development). Both families continue. The five properties used in Sections 3.2 and 3.3 — that the b cyclic k-mers of the block are distinct and each contain `X`; that (4), (6) are common superstrings of length $bt+k-1$; that each merge assembling (5), (7) is at overlap $k-1$; that the resulting strings are pairwise overlap-free; and that the block's edge set decomposes into the three pieces used — are all finite checks on the block itself, independent of t, and each was carried out. Granting the induction over t, which is bookkeeping rather than mathematics, one obtains

$$\rho_4 \;\geq\; \lim_{t\to\infty} \frac{19t}{10t+3} \;=\; \frac{19}{10} \;=\; 1.9, \qquad \rho_5 \;\geq\; \lim_{t\to\infty} \frac{25t}{13t+4} \;=\; \frac{25}{13} \;=\; 1.9230\ldots,$$

narrowing the brackets of Corollary 5 to [1.9, 2] and [25/13, 3]. These two statements are *not* part of the machine-checked development, which instantiates the families only at the six values of t tabulated above; they are recorded here as what the constructions give and are not used anywhere below.

4. String length four

Theorem 10. *Let $S_4^{(4)}$ be the instance of 40 strings of length 4 of Section 3.2. Then $\mathrm{OPT}(S_4^{(4)}) = 43$, and there is a greedy certificate for $S_4^{(4)}$ whose output has length 76. Consequently*

$$\rho_4 \;\geq\; \frac{76}{43} \;>\; \frac{7}{4} \;=\; 2 - \frac{1}{4},$$

so the Cazaux–Rivals conjecture fails at $k = 4$.

Proof sketch. The instance, the word $w_4^{(4)}$ of (4), the list of 39 merges and the output are explicit finite data. Three properties of them are decided by evaluation: that the 40 strings are pairwise distinct; that $w_4^{(4)}$ is a common superstring of the instance; and that the list of merges is a greedy certificate in the sense of Definition 6, that is, that at each of the 39 steps the two chosen live strings are distinct and their overlap equals the maximum over all ordered pairs of distinct live

strings. The last is the largest of the three computations, and it is exhaustive only over the pairs available at each step. For the optimum, no search is performed: $|w_4^{(4)}| = 43 = 40 + 4 - 1$, so Lemma 8 makes it shortest. Finally $7 \cdot 43 = 301 < 304 = 4 \cdot 76$. □

Theorem 11. *With $S_6^{(4)}$ (60 strings, optimum 63, a legal greedy output of length 114) and $S_{23}^{(4)}$ (230 strings, optimum 233, a legal greedy output of length 437),*

$$\rho_4 \;\geq\; \frac{114}{63} \;>\; \frac{9}{5} \qquad \textit{and} \qquad \rho_4 \;\geq\; \frac{437}{233} \;>\; \frac{15}{8}.$$

Proof sketch. Identical in shape to Theorem 10, with 59 and 229 merges respectively and with $|w_6^{(4)}| = 63 = 60+4-1$ and $|w_{23}^{(4)}| = 233 = 230+4-1$. The two arithmetic facts are $9 \cdot 63 = 567 < 570 = 5 \cdot 114$ and $15 \cdot 233 = 3495 < 3496 = 8 \cdot 437$. □

5. String length five

Theorem 12. *Let $S_5^{(5)}$ be the instance of 65 strings of length 5 of Section 3.3. Then $\mathrm{OPT}(S_5^{(5)}) = 69$, and there is a greedy certificate for $S_5^{(5)}$ whose output has length 125. Consequently*

$$\rho_5 \;\geq\; \frac{125}{69} \;>\; \frac{9}{5} \;=\; 2 - \frac{1}{5},$$

so the Cazaux–Rivals conjecture fails at $k = 5$.

Proof sketch. As for Theorem 10: distinctness of the 65 strings, the superstring property of $w_5^{(5)}$ and the legality of the 64 merges are decided by evaluation; $|w_5^{(5)}| = 69 = 65+5-1$ makes $w_5^{(5)}$ shortest by Lemma 8; and $9 \cdot 69 = 621 < 625 = 5 \cdot 125$. □

Theorem 13. *With $S_7^{(5)}$ (91 strings, optimum 95, a legal greedy output of length 175) and $S_{13}^{(5)}$ (169 strings, optimum 173, a legal greedy output of length 325),*

$$\rho_5 \;\geq\; \frac{175}{95} \;>\; \frac{11}{6} \;=\; 2 - \frac{1}{6} \qquad \textit{and} \qquad \rho_5 \;\geq\; \frac{325}{173} \;>\; \frac{15}{8}.$$

In particular the conjectured value $2 - 1/5$ is wrong at $k = 5$ by more than one step of the sequence $2 - 1/k$ itself.

Proof sketch. As above, with 90 and 168 merges; $11 \cdot 95 = 1045 < 1050 = 6 \cdot 175$ and $15 \cdot 173 = 2595 < 2600 = 8 \cdot 325$. □

6. Consequences

Proof of Corollary 3. For $k = 3$ the value $\rho_3 = 9/5 > 5/3$ is [1]; alternatively, Theorem 15(iii) below exhibits an instance of 3-SCS with $\rho_3 > 5/3 = 2 - 1/3$, which needs their construction but not their proof. For $k = 4$ and $k = 5$, Theorem 2(i), (ii). For $k = 6$, Theorem 15(iii) again, or [1]. For $k \geq 7$, $\rho_k \geq 2 > 2 - 1/k$ by [1]. □

Proof of Corollary 4. $\rho_4 \geq 114/63 > 9/5$ by Theorem 11, and $\rho_3 = 9/5$ by [1]. □

Corollary 4 is the qualitative statement we would single out. Before it, the sharpest published information at $k = 4$ was $1.75 \leq \rho_4 \leq 2$ (Remark 1), which is consistent with $\rho_4 = \rho_3$, and indeed consistent with ρ_k being constant on $3 \leq k \leq 5$; the two block families show instead that ρ_4 and ρ_5 both exceed $15/8$, hence both exceed $\rho_3 = 9/5$, and indeed both exceed the value $2 - 1/6$ that the conjecture

assigns to a longer string length. We have no upper bound to offer: whether $\rho_4 = 2$, and so whether the greedy conjecture is tight already at string length 4, remains open. The greedy conjecture itself is recorded in [4] as known for strings of length at most 3 and for strings of length *exactly* 4 — the mixed case of strings of length at most 4 is called "uncharted territory" there — while at $k = 5$ the best published upper bound is the generic $\rho_5 \leq 3$ of (2).

Remark 14. All lower bounds here, like those of [2] and [1], are statements about adversarial tie-breaking: they exhibit one legal execution, not the behaviour of every one. That this is the right reading of ρ_k is the standard convention and is the framing of [1]; that it is not vacuous in the present instances is Theorem 16(v), which exhibits a second legal execution on the $k = 6$ instance of [1] attaining the optimum exactly.

7. The published constructions, reproduced

The machinery of Section 2 was calibrated on the constructions of [1] before it was used for new ones. The following theorem is a reproduction and claims no novelty; every number in parts (i) and (ii) is printed in [1].

Theorem 15. (i) *(The $k = 6$ block of* [1]*.) The union of the cyclic* 6*-spectra of* `ABAABX` *and* `ABX` *consists of* 9 *distinct strings; the word* `ABAABXABXABAAB` *of length* $14 = 9 + 6 - 1$ *is a common superstring, hence shortest; there is a greedy certificate with output g of length* 18*; and the two words* $p_1 =$ `XABAABXABAA` *and* $p_2 =$ `BXABXABX`*, of lengths* 11 *and* 8*, satisfy* $\operatorname{ov}(p_2, p_1) = 1$ *and* $p_2 \odot p_1 = g$*.*

(ii) *(The $k = 3$ block of* [1]*.) The* 3*-spectrum of* `ABXAXAB` *consists of* 5 *distinct strings with optimum* 7*, and there is a greedy certificate with output of length* 9*.*

(iii) *Iterating those two blocks gives* $\rho_3 > 5/3$ *(an instance of* 30 *strings, optimum* 32*, greedy output* 54*) and* $\rho_6 > 11/6$ *(an instance of* 63 *strings, optimum* 68*, greedy output* 126*).*

Proof sketch. Each item is the same finite check as in Sections 4 and 5: distinctness, the superstring property of the exhibited word, the length identity $n+k-1$ feeding Lemma 8, and the replay of the certificate. For (i) the two piece lengths, their overlap and the identity $p_2 \odot p_1 = g$ are likewise evaluated. For (iii), $5 \cdot 32 = 160 < 162 = 3 \cdot 54$ and $11 \cdot 68 = 748 < 756 = 6 \cdot 126$. □

Part (iii) is what Corollary 3 uses at $k = 3$ and $k = 6$, so that only the range $k \geq 7$ of that corollary rests on a theorem quoted from elsewhere. Note that Theorem 15(iii) reproduces the *refutations* of [1] at those two string lengths, not their theorems: $\rho_3 = 9/5$ and $\rho_k \geq 2$ for all $k \geq 6$ are statements about infinite families and are not reproved here.

Theorem 16 (Controls)**.** (i) *Lemma 7 fails without distinctness: for the multiset* $(\mathtt{AA}, \mathtt{AA})$ *and* $w =$ `AA` *one has* $|w| = 2 < 3 = 2 + 2 - 1$*.*

(ii) *A list of merges that leaves more than one string live is rejected.*

(iii) *One block alone does not suffice at $k = 4$: $S_1^{(4)}$ has* 10 *strings, optimum* 13 *and a legal greedy output of length* 19*, and* $19/13 < 7/4$*. The iteration over t, not the block, is what crosses the threshold.*

(iv) *The certificate checker is not permissive: replacing the first merge of the certificate of Theorem 10 by a pair of overlap* 2*, rather than the maximum* 3*, makes it reject the whole certificate. Nor is the superstring checker constant-true: deleting the first letter of* $w_4^{(4)}$ *destroys the property.*

(v) *On the* $k = 6$ *instance of Theorem 15(i) there is a second legal greedy execution whose output has length* 14*, the optimum.*

Items (i)–(iv) are negative controls on the definitions rather than results: (i) shows that the distinctness hypothesis of Lemma 7 is load-bearing, (ii) and (iv) that a certificate can fail to be accepted — so an accepted certificate carries information — and (iii) that the theorems are not artefacts of a single block. Item (v) is the content of Remark 14.

8. How the blocks were found

This section is expository: nothing in it is used by any theorem above, and none of it is part of the machine-checked development. It is included because the search space of k-SCS instances is astronomically large and the two blocks of Section 3 did not come out of it by luck.

Remark 17 (the design principle; not part of the machine-checked development). Let $S \subseteq \Sigma^k$ and let $G(S)$ be the de Bruijn subgraph whose vertices are $(k-1)$-mers and whose edges are the elements of S, the string $s_0 \cdots s_{k-1}$ being an edge from $s_0 \cdots s_{k-2}$ to $s_1 \cdots s_{k-1}$. If $G(S)$ is balanced and weakly connected then $\mathrm{OPT}(S) = n + k - 1$ with $n = |S|$ [5]. By the endpoint-inheritance lemma of [6] — for live strings $x \neq y$ one has $\mathrm{ov}(x, y) = \mathrm{ov}(\mathrm{right}(x), \mathrm{left}(y)) \leq k - 1$, where $\mathrm{right}(x)$ is the last *input* string merged into x and $\mathrm{left}(y)$ the first of y, the bound because those are two distinct inputs of length k — greedy's merges at overlap $k-1$ concatenate edge-disjoint trails of $G(S)$, and this phase can stop only when no live pair $x \neq y$ has the right endpoint vertex of x equal to the left endpoint vertex of y. On a balanced graph that forces every live trail to be *closed* with pairwise distinct base vertices: each open trail contributes an out-excess at its start and an in-excess at its end, balance makes the multisets of starts and ends coincide, and the stopping condition then matches each trail to itself. A closed trail of L edges spells a string of length $L + k - 1$, so if the phase leaves c pieces and greedy then recovers total overlap O, the output has length $n + c(k-1) - O$ and

$$\frac{|g|}{\mathrm{OPT}(S)} \;=\; \frac{n + c(k-1) - O}{n + k - 1} \;\longrightarrow\; 1 + (k-1)\,\frac{c}{n} \qquad (n \to \infty,\ O/n \to 0).$$

Maximising greedy's ratio is minimising the average length of a closed trail. This observation is not new in substance — read in the opposite direction, bounding the number of trails in a cover bounds greedy from above, which is the engine of the $\rho_3 \leq 9/5$ proof of [1] — and the constraint that the average closed trail have length at least 2 is exactly the bound $\rho_k \leq (k+1)/2$ of [2]. Its use here is as a design specification: over cyclic block words B of length b whose k-mers are distinct and `X`-carrying, with some $(k-1)$-mer `X`-free (which is what connects the blocks), decompose the block's closed walk into c closed trails with pairwise overlap-free bases, and maximise $1 + c(k-1)/b$.

At $k = 4$ the block $B^{(4)}$ decomposes into the $c = 3$ closed trails carrying edges $\{9, 0, 8\}$, $\{2, 3, 7, 1\}$ and $\{6, 4, 5\}$, based at `XAA`, `XAB` and `XBA`, which spell the three

strings (5); $1+3\cdot 3/10 = 19/10$. At $k=5$ the block $B^{(5)}$ decomposes into the trails carrying $\{0,1,2,11,12\}$, $\{3,4,5,6,9,10\}$ and $\{7,8\}$, based at `XAAA`, `XAAB` and `XBXB`, spelling (7); $1+3\cdot 4/13 = 25/13$. The two-edge trail at $k=5$ is the 2-cycle on the 4-mers `XBXB` and `BXBX`, and it is available only because k is odd: the two vertices of a 2-cycle have period 2, so for even k a $(k-1)$-mer on such a cycle has equal first and last letters and cannot both begin with a fresh letter and end with a shared one — which is what the overlap-0 argument of Section 3.2 asks of a base. That is one concrete reason the $k=5$ optimum of the search below (25/13) exceeds the $k=4$ one (19/10), and a hint about where a $k=4$ upper bound might come from.

Remark 18 (a measurement; not part of the machine-checked development). We enumerated cyclic block words up to rotation and letter renaming, filtered by the three conditions of Remark 17, enumerated all simple cycles of the block graph, computed maximum-cardinality exact covers of the edge set by cycles, and tested every assignment of bases for the pairwise-overlap-0 condition; a second variant drops that condition and instead simulates greedy on the c pieces to measure the overlap O it recovers. Two calibrations came out exactly right: at $k=3$, over blocks of length at most 12 on two shared and one fresh letter, the maximum is $9/5$ and is attained only at $b=5$, by four blocks up to rotation and renaming, one of which is `ABXAX`, the block of [1] — and $9/5$ is the value they *prove* to be ρ_3; at $k=6$, $b=9$, the search returns the construction of [1] and its ratio $2 = 1+(2\cdot 5-1)/9$, including the one unit of overlap it loses. Eight sweeps were run in all, each over one alphabet and one range of block lengths:

k	alphabet	block lengths	threshold
3	2 shared, 1 fresh	$3 \le b \le 12$	ratio ≥ 1.8
4	2 shared, 1 fresh	$4 \le b \le 16$	ratio ≥ 1.9
4	3 shared, 1 fresh	$4 \le b \le 14$	ratio ≥ 1.9
4	2 shared, 2 fresh	$4 \le b \le 13$	ratio ≥ 1.9
4	2 shared, 1 fresh	$4 \le b \le 11$	ratio ≥ 1.9, overlap-0 dropped
5	2 shared, 1 fresh	$5 \le b \le 16$	ratio ≥ 1.8
5	2 shared, 1 fresh	$5 \le b \le 14$	ratio $\ge 25/13$, overlap-0 dropped
6	2 shared, 1 fresh	$6 \le b \le 14$	ratio > 2

No sweep used more than three shared or more than two fresh letters, and none combined three shared letters with two fresh ones. Nothing in these ranges beats 19/10 at $k=4$, 25/13 at $k=5$, or the value 2 of [1] at $k=6$. This is a measurement of a search, not an upper-bound proof: it examines families built from a single repeated block over small alphabets and short blocks. Widening it is not the productive direction — the number of block words grows exponentially in b — whereas an upper bound for $k=4$ in the style of the trail-decomposition proof of $\rho_3 \le 9/5$ in [1] would settle the cell; Remark 17 says exactly what such an argument has to bound, namely the density of closed trails in a balanced de Bruijn subgraph whose base vertices are pairwise overlap-free.

9. Verification

Lemmas 7 and 8 and Theorems 10, 11, 12, 13, 15 and 16 have been formalised and checked by the Lean 4 [11] kernel against Mathlib [10], with the following division of labour. Overlap, merge, occurrence as a contiguous block, the common-superstring

predicate, the legality test of Definition 6 and the replay of a certificate are defined as computable functions; that ov is the overlap and not merely a function of that name is itself proved, in the two halves "the length-ov(p, q) suffix of p is a prefix of q" and "no longer suffix is". Lemma 7 is proved from first principles and depends on no axiom beyond propositional extensionality and quotient soundness — in particular on no evaluation of compiled code — and Lemma 8 is an immediate consequence of it. This matters, because it is the half that could flatter the results. For each instance three finite facts — that the input strings are pairwise distinct, that the exhibited word is a common superstring, and that the list of merges replays as a legal execution — are discharged by compiled evaluation (Lean's `native_decide`), which trusts the compiler and the runtime for those three evaluations and records that trust as an axiom per evaluation; the largest of them replays 229 merges, each maximising the overlap over all ordered pairs of distinct live strings. No proof uses `sorry`; apart from the native-evaluation axioms just described, the only axioms occurring anywhere are Lean's standard `propext`, `Classical.choice` and `Quot.sound`; and there is no external certificate file. What is *not* formalised is: the passage from an instance to ρ_k (Section 2), since the development defines no supremum, and with it Theorem 2 and Corollaries 3–5 as stated; the results quoted from [1], [2] and [3] in those corollaries; the identification, in Section 3, of the instances as spectra of block words, the instances entering the development as explicit lists of strings; and everything in Remarks 9, 17 and 18.

Independently of the formal development, an implementation written against the definitions of [1] was used first as a calibration and then as a cross-check: it reproduces every number printed in [1] for the two constructions of Theorem 15, computes exact optima by dynamic programming over the overlap graph, simulates both block families for $t \leq 5$, and at $t = 1$ enumerates *all* legal greedy executions, the maximum output length being exactly the certified 19 at $k = 4$ and 25 at $k = 5$ — so on one block the exhibited execution is not merely legal but worst possible. That implementation also confirmed that along every certificate above no live string is ever a substring of another, so the executions are legal for the substring-eliminating formulation of Greedy as well. None of this is part of the machine-checked development. The repository is private: repository reference to be added at submission.

References

[1] N. Chukhin, A. S. Kulikov, I. Mihajlin and A. Smal, *The Greedy Superstring Algorithm Achieves Ratio 2 for Strings of Length 6 Already*, arXiv:2608.20018 (v1, 20 August 2026; v2, 24 August 2026), doi:10.48550/arXiv.2608.20018. All statements cited above are those of version v2, read from its LaTeX source; the quoted display (3) is from its introduction, the theorem $\rho_3 = 9/5$ from its sections on the lower and upper bounds for ρ_3, and the theorem $\rho_k \geq 2$ for $k \geq 6$ from its section on the lower bound for ρ_k. As of 30 August 2026 the arXiv record carries no journal reference and v2 is the current version.

[2] B. Cazaux and E. Rivals, *Relationship between superstring and compression measures: New insights on the greedy conjecture*, Discrete Applied Mathematics **245** (2018), 59–64, doi:10.1016/j.dam.2017.04.017. Proposition 2 is the lower bound $\rho_r \geq 2 - 1/r$, Theorem 3 the two-sided bound (2), and the conjecture is the sentence quoted in Section 1.2, immediately following Theorem 3; the sentence about [3] quoted in Remark 1 is from Section 1.

[3] A. S. Kulikov, S. Savinov and E. Sluzhaev, *Greedy Conjecture for Strings of Length 4*, in: Combinatorial Pattern Matching (CPM 2015), Lecture Notes in Computer Science **9133**, Springer, 2015, pp. 307–315, doi:10.1007/978-3-319-19929-0_26. Cited here only through its published abstract, quoted in Remark 1; the full text is closed access.

[4] M. S. Nikolaev, *Greedy Conjecture for the Shortest Common Superstring Problem and Its Strengthenings*, in: String Processing and Information Retrieval (SPIRE 2024), Lecture Notes in Computer Science **14899**, Springer, 2024, pp. 233–248, doi:10.1007/978-3-031-72200-4_18; arXiv:2407.20422. The two phrases quoted above are from the introduction of the arXiv version v1.

[5] P. A. Pevzner, *l-Tuple DNA Sequencing: Computer Analysis*, Journal of Biomolecular Structure and Dynamics **7** (1989), no. 1, 63–73, doi:10.1080/07391102.1989.10507752. Only the elementary half of the lemma attributed to this paper is used, and it is proved here (Lemma 7).

[6] A. Blum, T. Jiang, M. Li, J. Tromp and M. Yannakakis, *Linear approximation of shortest superstrings*, in: Proceedings of the 23rd Annual ACM Symposium on Theory of Computing (STOC 1991), ACM, 1991, pp. 328–336, doi:10.1145/103418.103455; journal version, Journal of the ACM **41** (1994), no. 4, 630–647, doi:10.1145/179812.179818. Quotations and lemma references above are to the journal version, whose Section 1 carries the sentence quoted in Section 1.1 and whose Section 2 carries, after its Claim 1, the endpoint-inheritance statement that the maximum overlap of two live strings s, t equals $\mathrm{ov}(\mathrm{last}(s), \mathrm{first}(t))$.

[7] J. A. Storer, *Data Compression: Methods and Theory*, Computer Science Press, Rockville, Md., 1988.

[8] J. Tarhio and E. Ukkonen, *A greedy approximation algorithm for constructing shortest common superstrings*, Theoretical Computer Science **57** (1988), no. 1, 131–145, doi:10.1016/0304-3975(88)90167-3.

[9] J. S. Turner, *Approximation algorithms for the shortest common superstring problem*, Information and Computation **83** (1989), no. 1, 1–20, doi:10.1016/0890-5401(89)90044-8.

[10] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, pp. 367–381, doi:10.1145/3372885.3373824.

[11] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE-28, Lecture Notes in Computer Science, Springer, 2021, pp. 625–635, doi:10.1007/978-3-030-79876-5_37.